中图法分类号:TP309　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)07-1787-27

论文引用格式:Zhou D W, Xu Y B, Wang N N, Liu D C, Peng C L and Gao X B. 2024. Generalized adversarial defense against unseen attacks: a survey. Journal of Image and Graphics, 29(07):1787-1813(周大为, 徐一搏, 王楠楠, 刘德成, 彭春蕾, 高新波. 2024. 针对未知攻击的泛化性对抗防御技术综述. 中国图象图形学报, 29(07):1787-1813)[DOI:10.11834/jig.230423]

针对未知攻击的泛化性对抗防御技术综述

周大为[1#],徐一搏[1#],王楠楠[1*],刘德成[1],彭春蕾[1],高新波[2]

1. 西安电子科技大学空天地一体化综合业务网全国重点实验室,西安　710071;
2. 重庆邮电大学重庆市图像认知重点实验室,重庆　400065

摘　要: 在计算机视觉领域,对抗样本是一种包含攻击者所精心设计的扰动的样本,该样本与其对应的自然样本的差异通常难以被人眼察觉,却极易导致深度学习模型输出错误结果。深度学习模型的这种脆弱性引起了社会各界的广泛关注,与之相对应的对抗防御技术得到了极大发展。然而,随着攻击技术和应用环境的不断发展变化,仅实现针对特定类型的对抗扰动的鲁棒性显然无法进一步满足深度学习模型的性能要求。由此,在尽可能不依赖对抗样本的情况下,通过更高效的训练方式和更少的训练次数,达到一次性防御任意种类的未知攻击的目标,是当下亟待解决的问题。期望所防御的未知攻击要有尽可能强的未知性,要在原理、性能上尽可能彻底地不同于训练阶段引入的攻击。为进一步了解未知攻击的对抗防御技术的发展现状,本文以上述防御目标为核心,对本领域的研究工作进行全面、系统的总结归纳。首先简要介绍了研究背景,对防御研究所面临的困难与挑战进行了简要说明。将未知对抗攻击的防御工作分为面向训练机制的方法和面向模型架构的方法。对于面向训练机制的方法,根据防御模型所涉及的最基本的训练框架,从对抗训练、自然训练以及对比学习3个角度阐述相关工作。对于面向模型架构的方法,根据模型结构的修改方式从目标模型结构优化、输入数据预处理两个角度分析相关研究。最后,分析了现有未知攻击防御机制的研究规律,同时介绍了其他相关的防御研究方向,揭示了未知攻击防御研究的整体发展趋势。不同于一般对抗防御综述,本文注重在未知性极强的攻击上的防御的调研与分析,对防御机制的泛化性、通用性提出了更高的要求,希望能为未来防御机制的研究提供更多有益的思考。

关键词: 对抗防御;未知对抗攻击;对抗训练;数据预处理;深度学习

Generalized adversarial defense against unseen attacks: a survey

Zhou Dawei[1#], Xu Yibo[1#], Wang Nannan[1*], Liu Decheng[1], Peng Chunlei[1], Gao Xinbo[2]

1. State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an　710071, China;
2. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing　400065, China

Abstract: Deep learning-based models have achieved impressive breakthroughs in various areas in recent years. However, they are vulnerable when their inputs are affected by imperceptible but adversarial noises, which can easily lead to wrong outputs. To tackle this problem, many defense methods have been proposed to mitigate the effect from these threat models for deep neural networks. As adversaries seek to improve the technologies of disrupting the models' performances,

收稿日期:2023-06-30;**修回日期**:2024-03-11;**预印本日期**:2024-03-18

#共同一作; * **通信作者**:王楠楠　nnwang@xidian.edu.cn

基金项目:国家自然科学基金项目(U22A2096, 62036007, 62306227);陕西省自然科学基础研究计划(2022JQ-696)

Supported by: National Natural Science Foundation of China (U22A2096, 62036007, 62306227); Natural Science Basic Research Plan in Shaanxi Province of China (2022JQ-696)

an increasing number of attacks that are unseen to the model during the training process are emerging. Thus, the defense mechanism, which defends against only some specific types of adversarial perturbations, is becoming less robust. The ability of a model to generally defend against various unseen attacks becomes pivotal. Unseen attacks should be as different as possible from the attacks used in the training process in terms of theory and attack performance rather than adjustment of parameters from the same attack method. The core is to defend against any attacks via efficient training procedures, while the defense is expected to be as independent as possible from adversarial attacks during training. Our survey aims to summarize and analyze the existing adversarial defense methods against unseen adversarial attacks. We first briefly review the background of defending against unseen attacks. One of the main reasons that the model is robust against unseen attacks is that it can extract robust features through a specially designed training mechanism without explicitly designing a defense mechanism that has special internal structures. A robust model can be achieved by modifying its structure or designing additional modules. Therefore, we divide these methods into two categories: training mechanism-based defense and model structure-based defense. The former mainly seeks to improve the quality of the robust feature extracted by the model via its training process. 1) Adversarial training is one of the most effective adversarial defense strategies, but it can easily overfit to some specific types of adversarial noises. Well-designed attacks for training can explicitly improve the model's ability to explore the perturbation space during training, which directly helps the model learn more representative features compared with traditional adversarial attacks in the perturbation space. Adding regularization terms is another way to obtain robust models by improving the robust features from the basic training process. Furthermore, we introduce some adversarial training-based methods combined with knowledge from other domains, such as domain adaptation, pre-training, and fine tuning. Different examples make different contributions to the model's robustness. Thus, example reweighting is also a way to achieve robustness against attacks. 2) Standard training is the most basic training method in deep learning. Data augmentation methods focus on example diversity of standard training, while adding regularization terms into standard training aims to enhance the model outputs' stabilization. Pre-training strategy aims to achieve a robust model within a predefined perturbation bound. 3) We also found that contrastive learning is a useful strategy as its core ideas about feature similarity match well with the goal of acquiring representative robust features. Model structure-based defense, meanwhile, mainly focuses on intrinsic drawbacks from the model's structure. It is divided into structure optimization for target network methods and input data pre-processing methods according to how the structures are modified. 1) Structure optimization for target network aims to enhance the model's ability to obtain useful information from inputs and features because the network itself is susceptible to variations from them. 2) Input data pre-processing focuses on eliminating the threats from examples before feeding them into the target network. Removing adversarial noise from inputs or detecting adversarial examples to reject them are two popular strategies because they are easily modeled and rely less on adversarial training examples compared with other methods such as adversarial training. Finally, we analyze the trends of research in this area and summarize some research on other related domains. 1) Defending against multiple adversarial perturbation well cannot make sure that the model is robust against various unseen attacks but contributes to the improvement of robustness against one specific type of perturbation. 2) With the development of defense against unseen adversarial attacks, some auxiliary tools such as the accelerating module have been proposed. 3) Defense against unseen common corruptions is beneficial for applications of defense methods because adversarial perturbations cannot represent the whole perturbation space in the real world. To summarize, defending against attacks that are totally different from the attacks during training has stronger generalizability. The analysis based on this goal shows differences from traditional surveys about adversarial defense. We hope that this survey can further motivate research on defending against unseen adversarial attacks.

Key words: adversarial defense; unseen adversarial attacks; adversarial training; data pre-processing; deep learning

0 引　言

随着信息技术的不断进步，以深度学习为核心的人工智能技术展现出令人印象深刻的强大性能，深度学习因此得到广泛关注，与之相关的图像分类（Krizhevsky 等，2017；He 等，2019）及其在半监督学习上的扩展（吕昊远 等，2021）、遥感影像处理（王鑫

等，2019）、语音识别（Abdel-Hamid 等，2014；戴礼荣等，2017）与合成（Weng 等，2023）、目标检测（Ren 等，2015；李科岑 等，2022；袁珑 等，2022）等飞速发展，并在人类社会引发了广泛而深刻的变革。然而，面对日益复杂多变的应用环境，深度神经网络的安全问题也逐渐暴露出来。例如，在计算机视觉领域，攻击者在交通信号牌上恶意添加噪声导致自动驾驶系统无法正确判断信号内容，做出不符合实际路况的危险判断，引发行人碰撞等安全事故（Modas 等，2020）；人脸识别系统在识别添加特定扰动的人脸时会输出错误的识别结果（Zheng 等，2023；Li 等，2023b），导致个人身份、单位财务等机密信息的泄露等等。深度学习的这种缺陷为智能产品落地、数据隐私保护、公共环境安全等带来巨大阻碍，引发了一系列人工智能伦理危机、信任危机（张钹 等，2020），带来了一系列社会、经济和文化等问题，极大地限制了相关技术的进一步发展。

在一系列的深度学习安全问题中，对抗扰动问题尤为突出。Szegedy 等人（2014）开创性地发现，在输入中添加精心设计但难以被人肉眼察觉的噪声，可导致深度学习模型以较高的置信度输出错误结果。深度学习的这种脆弱性受到社会各界广泛关注，与此相关的对抗攻击技术不断发展。根据攻击者是否可以访问模型内部信息，可将现有对抗攻击分为两种，一种是白盒攻击（Moosavi-Dezfooli 等，2016；Liu 等，2022；Agnihotri 等，2023），即攻击者可以利用模型内部信息制作扰动；另一种是黑盒攻击（Sun 等，2022；Williams 和 Li，2023；Yin 等，2024），即攻击者在制作扰动时无法利用任何模型信息。前者对模型的破坏相对更大，而后者更符合实际情况，因为实际部署的模型通常都有一定的保护机制。对抗扰动严重损害了深度神经网络的性能，对深度学习系统构成了极大威胁。

为消除这种扰动带来的威胁，研究者们提出了多种对抗防御方法。其中，对抗训练被证明是最有效的防御方法之一，它通过在训练迭代时不断将对抗样本加入到训练集中，期望目标网络在训练阶段学习到对抗样本的鲁棒特征（Zhang 等，2019；de Jorge Aranda 等，2022；Wang 等，2023）。为了提升对抗训练的效率，FGSM（fast gradient sign method）（Goodfellow 等，2015）、PGD（projected gradient descent）（Madry 等，2018）等梯度攻击被引入训练过程中以生成对抗样本。预处理防御旨在设计一个去噪器，先将对抗样本输入去噪器得到去噪图像，再将去噪图像输入目标网络以输出预测结果，在不改变目标网络结构和参数的前提下提升模型鲁棒性（Jin 等，2019；Liao 等，2018；Yoon 等，2021）。另外，对抗样本检测旨在检测出对抗样本的存在，避免对抗样本被输入至目标网络中（Wang 和 Gong，2022）。除了以上方法，防御对抗扰动的方法还包括数据随机化等，它们从不同角度提升了模型鲁棒性。

然而，现有的防御方法面临鲁棒泛化性不佳的问题。例如，对抗训练很有可能在特定类型的扰动上陷入过拟合，从而导致模型在测试阶段的鲁棒性相对于训练阶段大幅下降，而即便这种问题得到缓解，这种防御通常也只能实现对单一扰动的鲁棒性（Rice 等，2020；Chen 等，2021）。攻击技术的不断发展为防御技术带来了多种多样的威胁，仅对特定的攻击实现鲁棒性未必能保证对其他类型的攻击也实现鲁棒性。为此，近期有部分工作追求同时实现对多种攻击的鲁棒性。然而，由于训练阶段引入了较多的攻击，这样的防御训练开销极大。在现实环境中，面临不断发展丰富的攻击手段，同时针对多种类型的扰动的防御未必能够学习到更具代表性的对抗扰动特征，仅在训练阶段引入多种攻击很难保证新的更强的攻击无法突破现有防御机制，从而无法有效防御这些未知攻击。对于防御机制而言，未知攻击显然威胁更大，更符合实际部署的环境特点，因此有必要用特殊方法来进一步设计、改进防御机制以处理这种更广泛的威胁。

目前的有关对抗防御的综述没有充分关注未知攻击的泛化性防御问题。Akhtar 和 Mian（2018）从修改数据或训练机制、修改网络结构和添加网络模块 3 个角度系统介绍了现有的对抗防御方法。Liang 等人（2022）从模型优化、数据优化和额外网络模块 3 个角度分析总结了现有的防御机制，不同于该综述的是，上述综述对通用扰动的防御以及对抗样本检测等工作做了更加细致的归纳。Li 等人（2022）介绍了现有防御方法中的对抗训练、随机化、去噪、检测等方法。调研发现，现有的综述很好地归纳了对抗防御工作，然而对未知攻击的防御问题没有进行充分讨论。不同于现有综述，本文从未知攻击的角度出发，系统调研、归纳了近期计算机视觉领域针对未知对抗攻击的泛化性防御工作。这里，强调防御

完全不同于训练时的攻击，期望模型通过尽可能高效的训练方法获得对任意未知攻击的鲁棒性。因此，相比于传统的对抗训练综述，本文对防御机制的泛化性具有更高的关注度。

对抗防御模型的整体性能主要以训练方法和模型架构为基础。通过调研发现，防御模型在训练时所学习到的鲁棒特征是否在扰动空间具有足够的代表性是模型对不可见攻击的防御能力的关键。此外，考虑到深度神经网络本身的脆弱性，模型结构本身总是存在鲁棒性缺陷，由此添加新的模块、修改网络结构为实现未知攻击的泛化提供了可能，成为众多研究考虑的方向。根据泛化能力来源于内部模型结构的鲁棒性还是来自外部训练机制的鲁棒性，可将防御未知攻击的工作主要分为两种，一种是面向训练机制的防御，该策略旨在通过优化训练过程来学习更具代表性的对抗样本特征；另一种是面向模型架构的防御，该策略倾向于通过添加特殊组件、修改模型结构等方式弥补模型本身自有的鲁棒泛化性缺陷，以防御不可见噪声。图1展示了面向未知攻击的泛化性对抗防御的框架，相关工作在表1做了归纳。此外，本文还分析了针对未知常规扰动的泛化性防御，以及多类型扰动鲁棒性的相关工作、相关的基准数据集等。最后，本文对未知攻击的泛化性防御研究进行总结与展望。

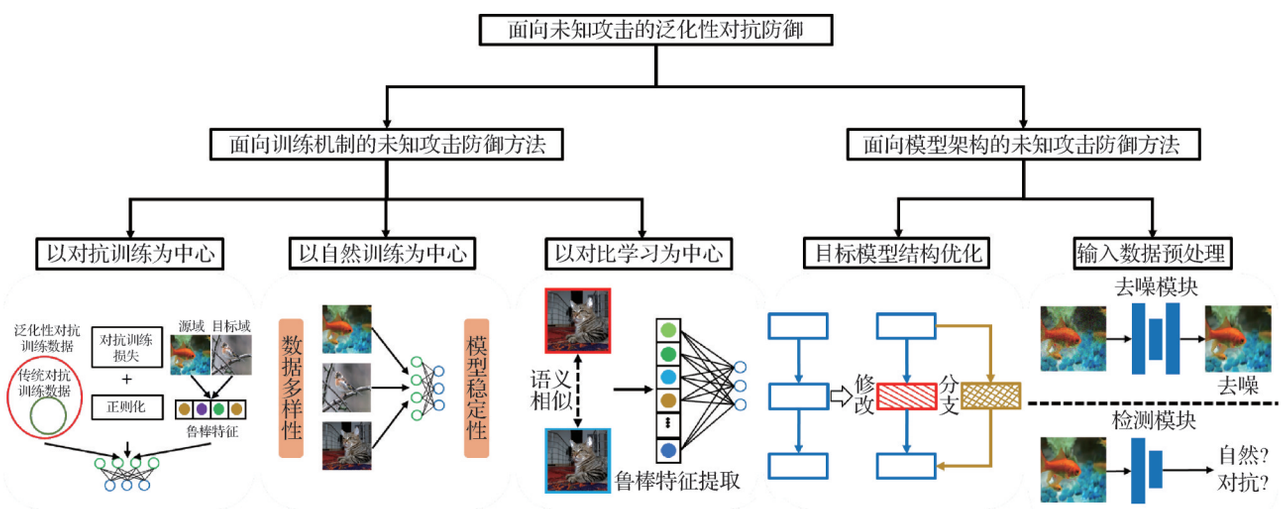

图1　面向未知攻击的泛化性对抗防御框架图

Fig. 1　The framework of the generalized defense against unseen attacks

1　面向训练机制的未知攻击防御方法

1.1　以对抗训练为中心

1.1.1　基于特殊攻击发掘潜在困难样本

传统的对抗训练缺乏对扰动空间的充分探索，这与模型在训练时使用的攻击方法有直接联系。为了解决这一问题，许多方法为对抗训练设计了更为特殊的攻击方式，力求突破传统对抗攻击所覆盖的扰动空间，学习更通用的鲁棒特征。

1）ADT。对抗扰动空间本身代表一种数据分布，已知的很多对抗训练方法仅用某种特定的攻击探索这个分布的某个局部区域，缺乏全局性的探索。Dong等人(2020)由此提出了ADT(adversarial distributional training)，通过对每个样本周围的扰动空间进行整体扰动来探索潜在的对抗样本，将其用于对抗训练，具体为

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\max_{p(\delta_i)}E_{p(\delta_i)}\left[L\left(f\left(\boldsymbol{x}_i+\boldsymbol{\delta}_i\right),\boldsymbol{y}_i\right)\right] \tag{1}$$

式中，$p(\boldsymbol{\delta}_i)$代表了样本$\boldsymbol{x}_i$附近的扰动分布，$\boldsymbol{y}_i$是其真实标签，f是目标分类网络，N是样本总数。通过该对抗训练对损失函数L的优化，目标网络可学到每个样本周围更加丰富的对抗样本特征，有效探索扰动空间，从而泛化到不可见攻击。这种数据分布扰动对扰动空间的探索在理论上更直接、范围更广，但其探索能力依赖于对这种分布的建模方法。

2）DMAT。DMAT(dual manifold adversarial training)(Lin等，2020)是一种基于生成思想的防御，具体为

$$\min_{\theta}\sum_{i}\left(A_i+B_i\right)$$
$$A_i=\max_{\boldsymbol{\delta}\in\Delta}L\left(f_\theta\left(g\left(\boldsymbol{z}_i\right)+\boldsymbol{\delta}\right),\boldsymbol{y}_i\right)$$
$$B_i=\max_{\boldsymbol{\eta}\in\Lambda}L\left(f_\theta\left(g\left(\boldsymbol{z}_i+\boldsymbol{\eta}\right)\right),\boldsymbol{y}_i\right)\tag{2}$$

式中，$\boldsymbol{\delta}$和$\boldsymbol{\eta}$分别是实例层和特征层的扰动，各自的扰动范围分别是Δ和Λ，L是损失函数，$\boldsymbol{y}_i$是特征$\boldsymbol{z}_i$的真实标签，f_θ是损失函数。特征层的扰动输入生成器g后得到的便是分布在一个特定的流形空间上的对抗样本，整体目标便是同时学习在和不在该特定流形上的对抗样本的特征，实现鲁棒防御。

除了生成器本身可提升样本多样性外，DMAT相关实验说明，仅在非流形对抗样本上训练未必能泛化到流形对抗样本，仅在流形对抗样本上训练也未必能泛化到非流形对抗样本，因此要实现更泛化的防御，就必须要同时处理它们。但是，这样的方法对这种流形做出了特定的设定，其性能依赖于生成器的质量。

3）CCAT。Stutz等人（2020）从置信度出发提出了CCAT（confidence-calibrated adversarial training）方法。传统对抗训练偏向于使模型在学习过程中为对抗样本赋予较高的置信度，从而导致模型陷入对指定类型扰动的过拟合。CCAT利用了一个特殊的攻击方式，即对于一个给定样本，最大化除正确标签外的其他所有标签的置信度，具体为

$$\max_{\|\boldsymbol{\delta}\|\leqslant\varepsilon}\max_{i\neq y}f_i\left(\boldsymbol{x}+\boldsymbol{\delta};\theta\right)\tag{3}$$

式中，$\boldsymbol{\delta}$和$\boldsymbol{\varepsilon}$分别是样本$\boldsymbol{x}$的扰动和对应最大扰动范围，f是分类器在第i个类别上的输出，θ是其参数，y是真实类别。与此不同的是，传统攻击方法是最小化正确类别的置信度，而该攻击是最大化除正确类别以外其他所有类的置信度，是非目标攻击。CCAT中还设计有标签平滑方式，具体为

$$\begin{cases}y_{\text{new}}=\alpha(\boldsymbol{\delta})\,one\_hot\left(y\right)+\left(1-\alpha(\boldsymbol{\delta})\right)\dfrac{1}{K}\\ \alpha(\boldsymbol{\delta})=\left(1-\min\left(1,\dfrac{\|\boldsymbol{\delta}\|}{\boldsymbol{\varepsilon}}\right)\right)^{\rho}\end{cases}\tag{4}$$

式中，*one_hot*表示将标签转化为one_hot格式向量，K是类别数量，y_{new}和y分别是平滑标签和原始标签，$\alpha(\boldsymbol{\delta})$是控制标签平滑程度的系数，平滑速率由$\rho$控制。由此可见，对抗样本距离原始样本越远，其标签越平滑。

由于CCAT对不同扰动程度的样本的真实标签进行了显式区分，测试时需引入一个置信度阈值，拒绝对不符合要求的样本（对抗样本）进行测试。对抗训练在产生对指定范围内的样本的鲁棒性时，会对超出该范围的对抗样本赋予低置信度。CCAT在训练中主动降低对抗样本的置信度，从而使最终模型具备延伸到训练扰动范围外的预测能力，由此在测试时，通过阈值筛选，超出扰动范围的样本同样可被拒绝，这种方式使其能有效防御未知攻击。

4）PAT。PAT（perceptual adversarial training）（Laidlaw等，2021）方法创新性地对扰动的感知特性进行了深入探索。人类难以感知的对抗扰动通常很难通过精确的数学模型来描述，为此许多防御算法将所防御的扰动直接限定在一个指定的范数范围内，然而这种度量方式并不完全符合人类感知的范围。考虑到深度神经网络自身强大的拟合能力，PAT调用了基于神经网络的距离度量方式LPIPS（learned perceptual image patch similarity）（Zhang等，2018），设计了新的攻击方式NPTM（neural perceptual threat model）。LPIPS的定义为

$$\text{dis}\left(\boldsymbol{x}_1,\boldsymbol{x}_2\right)=\left\|\boldsymbol{\phi}\left(\boldsymbol{x}_1\right)-\boldsymbol{\phi}\left(\boldsymbol{x}_2\right)\right\|\tag{5}$$

式中，$\boldsymbol{\phi}(\boldsymbol{x})$是经过处理后的样本在特定分类器$\boldsymbol{g}$上的特征层输出。假设$\boldsymbol{g}$有$k$层，对于给定的样本$\boldsymbol{x}$，每层经过归一化后的输出是$\boldsymbol{g}_i$，最终将每层的对应输出展开为一个向量作为$\boldsymbol{\phi}(\boldsymbol{x})$的输出。LPIPS距离充分利用了模型的内部信息，该度量方式被证明非常符合人眼感知范围。PAT利用LPIPS设计了如下攻击，具体为

$$f\left(\boldsymbol{x}^{\text{adv}}\right)\neq\boldsymbol{y}^{\text{true}}$$
$$\text{s.t. dis}\left(\boldsymbol{x},\boldsymbol{x}^{\text{adv}}\right)=\left\|\boldsymbol{\phi}\left(\boldsymbol{x}\right)-\boldsymbol{\phi}\left(\boldsymbol{x}^{\text{adv}}\right)\right\|_2\leqslant\boldsymbol{\varepsilon}\tag{6}$$

式中，$\boldsymbol{x}$和$\boldsymbol{x}^{\text{adv}}$分别为自然样本和对应的对抗样本，$\boldsymbol{y}^{\text{true}}$是真实标签，$f$是目标网络，$\boldsymbol{\varepsilon}$是最大扰动值。PAT的基本形式与标准对抗训练相似，唯一不同的是PAT在生成对抗样本时采用了基于LPIPS的距离度量方式。由于LPIPS是基于模型本身的度量方式，因此根据目标网络结构与计算LPIPS的网络结构是否相同，可将该对抗训练分为两种：二者结构不同时，调用一个预训练的模型计算该距离；二者结构相同时，LPIPS的距离会随着对抗训练过程中网络参数的优化而不断变化。

表1　面向未知攻击的防御方法归纳

Table 1　Summary of defense methods against unseen attacks

分类			方法
面向训练机制的未知攻击防御方法	以对抗训练为中心	基于特殊攻击发掘潜在困难样本	ADT（Dong 等，2020）、DMAT（Lin 等，2020）、CCAT（Stutz 等，2020）、PAT（Laidlaw 等，2021）、Lagrangian AT（Azizmalayeri 和 Rohban，2023）、MNG-AC（Madaan 等，2021）、IJSAT（Lau 等，2023）、GAT（Poursaeed 等，2021）、NCAT（Sriramanan 等，2022）、REx（Ibrahim 等，2022）、Semantic-GAT（Hsiung 等，2023）、SSAT（Jiao 等，2022）、RiFT（Zhu 等，2023）
		基于正则化发掘攻击语义信息	SRD（Silva 等，2022）、Consistency Regularization（Tack 等，2022）
		基于领域适应提取鲁棒特征	DIAL（Levi 等，2021）、AFD（Bashivan 等，2021）、ADV-4-ADV（Zheng 等，2024）
		基于预训练和微调实现高效鲁棒泛化	Pre-training to fine-tuning（Chen 等，2020b）、RiFT（Zhu 等，2023）
		基于样本重加权区分样本脆弱性	LRAT（Gao 等，2021）
	以自然训练为中心	基于数据增强提升样本多样性	RBF-CNN（Nandy 等，2020）、ALAT（Ho 等，2022）
		基于正则化约束模型训练	Gradreg（Yu 等，2018）、MMR-Universal（Croce 和 Hein，2019）、Manifold Regularization（Jin 和 Rinard，2020）、IART（Boopathy 等，2020）
		基于预训练对抗难以感知的攻击	VIB（Chhabra 等，2021）
	以对比学习为中心	基于无监督方法发掘鲁棒特征	RoCL（Kim 等，2020）、SwARo（Wahed 等，2022）、ACL（Jiang 等，2020）、Reverse Attack（Mao 等，2021）
		基于有监督方法学习鲁棒决策边界	TLA（Mao 等，2019）
面向模型架构的未知攻击防御方法	目标模型结构优化	基于特殊组件弥补模型特征缺陷	BPN（Wen 等，2020）、Selective Feature Regeneration（Borkar 等，2020）
		基于模型分支区分输入分布	ABS（Schott 等，2019）、DDC-AT（Xu 等，2021）
	输入数据预处理	对抗噪声去除	Defense-GAN（Samangouei 等，2018）、Feature Denoising（Xie 等，2019）、Online Alternate Generator（Li 等，2020）、CD-VAE（Yang 等，2021b）、Defense Transformer（Li 等，2023a）、JATP（Zhou 等，2021b）、ARN（Zhou 等，2021a）、CAFD（Zhou 等，2021c）、CDD-RED（Gong 等，2022）、Agnostic-Diffusion Model（Blau 等，2022）、DiffPure（Nie 等，2022）、Pixel and Feature Distribution Alignment（Xu 等，2022）
		对抗样本检测	I-Defender（Zheng 和 Hong，2018）、The Odds are Odd（Roth 等，2019）、NNIF（Cohen 等，2020）、LNG（Abusnaina 等，2021）、SimCat（Moayeri 和 Feizi，2021）、RSA（Drenkow 等，2022）、MIAED（Gao 等，2023）、DRAM（Tsai 等，2023）

5）Lagrangian AT。Azizmalayeri 和 Rohban（2023）在传统的攻击算法中添加了一个扰动的惩罚项，提出了如下的攻击方式并将其用于对抗训练中，具体为

$$\max_{\delta}\left[L\left(f_\theta\left(\boldsymbol{x}_i+\boldsymbol{\delta}\right),\boldsymbol{y}_i\right)-\lambda\left\|\boldsymbol{\delta}\right\|_2\right] \tag{7}$$

式中，L指目标网络f_θ的损失函数，$\boldsymbol{y}_i$是自然样本$\boldsymbol{x}_i$的真实标签，$\boldsymbol{\delta}$是添加进输入的扰动，λ是一个超参数。实际实现该攻击时，随着攻击的迭代，λ不断增加，动态调整算法对$\boldsymbol{\delta}$的大小的关注程度，而迭代步长不断减小，控制扰动值不要太大。

由上可知，相比于Lagrangian AT方法，PAT方法存在明显的缺陷。LPIPS是基于深度神经网络内层

的特征值来计算扰动距离的，为了输出正确的预测值，这些内层特征会对图像的不同区域施加不同的注意力，因而产生对不同图像区域的不同敏感性。当在敏感区域施加扰动时，内层特征变化剧烈，基于LPIPS距离的攻击因此会偏好在不敏感的区域施加扰动，保证对抗样本的扰动不超出指定范围。显然，在这种具有偏好的对抗样本上训练并不利于模型对未知攻击的防御，而Lagrangian AT方法不存在这种偏好问题，因而具有更强的泛化性。

此外，这种形式的攻击与CW（Carlini&Wagner）攻击（Carlini和Wagner，2017）非常相似，不同的是该攻击没有对λ进行像CW一样的精细优化，能够有效减少运算开销。相比于CW，该方法得到的扰动平均而言不会过小，探索到的对抗样本特征更加丰富，这对于未知攻击的泛化是有益的，因为并非所有攻击都以最小化扰动值为目标。

6）MNG-AC。要同时实现对多种类型的扰动的鲁棒性，一个直接的困难是如何降低可见攻击带来的太多运算开销。为此，Madaan等人（2021）提出在对抗训练过程中每轮只选定一种攻击进行训练，避免同时引入所有攻击增大运算负担。MNG-AC（meta-noise generator with adversarial consistency loss）在引入这种对抗训练的同时加入了如下的正则化项，具体为

$$\begin{cases} L_c = \dfrac{1}{3}\left(KL(p_{\mathrm{na}} \| q) + KL(p_{\mathrm{ad}} \| q) + KL(p_{\mathrm{au}} \| q)\right) \\ q = \dfrac{p_{\mathrm{na}} + p_{\mathrm{ad}} + p_{\mathrm{au}}}{3} \end{cases} \tag{8}$$

式中，KL代表KL（Kullback-Leibler）散度，p_{na}，p_{ad}，p_{au}分别是自然样本、对抗样本和经过数据增强的样本的预测概率。这种在多变化下的输出一致性和不同攻击下的鲁棒学习使得MNG-AC学习到高质量的不变特征，从而具备了泛化到不可见攻击的能力。

7）IJSAT。IJSAT（interpolated joint space adversarial training）（Lau等，2023）同时引入了插值和流形策略。这两种策略对于提升样本多样性，缓解过拟合等都很有帮助。IJSAT的优化目标为

$$\begin{cases} \min\limits_{\theta} \left\{ \max\limits_{\delta \in \Delta, \eta \in \Lambda} L_{\alpha}^{\mathrm{mixup}}\left(G\left(G^{-1}(\mathrm{mix}(\boldsymbol{x})) + \boldsymbol{\eta}\right) + \boldsymbol{\delta}\right) \right\} \\ \mathrm{mix}(\boldsymbol{x}) = \alpha \boldsymbol{x}_i + (1-\alpha)\boldsymbol{x}_j \\ L_{\alpha}^{\mathrm{mixup}}\left(\mathrm{mix}(\boldsymbol{x})\right) = \alpha L\left(f\left(\mathrm{mix}(\boldsymbol{x})\right), \boldsymbol{y}_i\right) + \\ \qquad (1-\alpha) L\left(f\left(\mathrm{mix}(\boldsymbol{x})\right), \boldsymbol{y}_j\right) \end{cases} \tag{9}$$

式中，L和$L_{\alpha}^{\mathrm{mixup}}$分别是插值前后的损失函数，$\alpha \in (0,1)$，$\boldsymbol{y}_i$、$\boldsymbol{y}_j$是对应标签。用插值法得到两个原始样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的混合样本$\mathrm{mix}(\boldsymbol{x})$，借助生成器$G$，该方法同时在特征层和样本上添加扰动$\boldsymbol{\eta}$和$\boldsymbol{\delta}$，并将这二者分别限定在$\Lambda$和$\Delta$范围内。插值法本身可以提升样本多样性，加之对抗样本在特征层面和实例层面均有扰动信息，因此该对抗训练流程可以充分探索扰动空间，从而泛化到未知攻击。

8）GAT。Poursaeed等人（2021）注意到，传统攻击只在自然样本上添加细微改动，相比于更高水平的风格变化，这种扰动是非常低阶的信息变化。如果只在这样低阶的对抗样本上进行训练，很可能导致模型无法泛化到更高阶的攻击。该方法通过在噪声和风格两个角度制作对抗扰动来生成语义信息变化更加丰富的对抗训练样本集，具体为

$$\begin{cases} \boldsymbol{\delta}_{\mathrm{style}}^{i+1} = \boldsymbol{\delta}_{\mathrm{style}}^{i} + \\ \quad \alpha \mathrm{sign}\left(\nabla_{\delta_{\mathrm{style}}^{i}} L\left(F\left(g\left(\boldsymbol{\delta}_{\mathrm{style}}^{i}, \boldsymbol{\delta}_{\mathrm{noise}}^{i}\right)\right), \boldsymbol{y}^{\mathrm{true}}\right)\right) \\ \boldsymbol{\delta}_{\mathrm{noise}}^{i+1} = \boldsymbol{\delta}_{\mathrm{noise}}^{i} + \\ \quad \beta \mathrm{sign}\left(\nabla_{\delta_{\mathrm{noise}}^{i}} L\left(F\left(g\left(\boldsymbol{\delta}_{\mathrm{style}}^{i}, \boldsymbol{\delta}_{\mathrm{noise}}^{i}\right)\right), \boldsymbol{y}^{\mathrm{true}}\right)\right) \end{cases} \tag{10}$$

式中，$\boldsymbol{\delta}_{\mathrm{noise}}$和$\boldsymbol{\delta}_{\mathrm{style}}$分别是高斯噪声和风格噪声，$g$是生成器，$F$是分类器，$L$是分类损失，$\boldsymbol{y}^{\mathrm{true}}$是正确类别，$\alpha$和$\beta$是攻击步长。生成器同时以高斯噪声和风格噪声作为输入来合成图像，可以只优化细微噪声或风格噪声，也可二者同时优化。该攻击类似于传统梯度攻击的迭代方式，所不同的是，该攻击方法将风格变化融入其中，且没有对扰动值施加范围限制，实现了对梯度攻击的高阶拓展。将这种高阶攻击生成的对抗样本作为训练集引入对抗训练中，由此便完成GAT（generative adversarial training）的整体流程。该方法可以选用模型不同层来生成对抗样本，这种方式充分把握了样本的语义信息，从风格角度极大地拓展了可见对抗样本特征在扰动空间的覆盖范围，而并没有在训练中显式地采用任何一种传统攻击方法。

9）NCAT。为了缓解在1范数的鲁棒训练的过拟合问题，Sriramanan等人（2021）设计了NuAT（nuclear-norm adversarial training）方法。但是，由于训练阶段的扰动不够多样，这种方法仍然无法有效处理过拟合。Sriramanan等人（2022）由此提出了一个扰动值的动态调整策略，即在训练时使扰动值从小到大线性变化，同时在训练时引入不同种类范数

的攻击，充分提升样本多样性，最终实现了跨范数的未知攻击防御NCAT（nuclear curriculum adversarial training）。

10）REx。不同的攻击、同种攻击的不同设置均能产生不同的对抗样本数据分布，这与领域泛化中的不同领域分布差异在理论上高度相似，成为在防御工作中引入领域泛化的基本动机。然而，将领域泛化引入对抗训练存在很大挑战。在标准的领域泛化中，不同领域的数据分布差异非常稳定，但在对抗训练中，这种差异并不稳定，因为随着训练的进行，模型参数不断变化，攻击效果也在不断变化。Ibrahim等人（2022）提出将REx（risk extrapolation）（Krueger等，2021）作为一个正则化项引入对抗训练中，利用领域泛化来提升鲁棒性。选用REx是因为它结构直观、计算量小、便于部署，有利于模型在训练时探索并处理更多的不同扰动分布的偏移。注意，该方法在训练阶段要引入多种类型的扰动以形成新的攻击方式，以更好地与领域泛化方法结合，因此正则化项在训练中必须与这种特殊攻击相互配合。

11）Semantic-GAT。现有防御工作对无穷范数攻击的研究较多，然而这种攻击不完全符合现实环境的干扰特点，旋转、对比度等干扰是实际环境中模型容易遇到的扰动，只防御无穷范数攻击的模型很难泛化到这种语义扰动。Hsiung等人（2023）由此提出了Semantic-GAT（generalized adversarial training to composite semantic perturbations）。为探索这些扰动，该方法首先对语义对抗样本训练集进行了构建。通常情况下，语义扰动空间是连续空间，并非离散空间，因此在该流程中语义攻击的设计借助了类似PGD（projected gradient descent）（Madry等，2018）的迭代梯度下降方法，具体为

$$\boldsymbol{\delta}_k^{i+1}=\boldsymbol{\delta}_k^i+\alpha\operatorname{sign}\left(\nabla_{\delta_k^i}L\left(F\left(A_k\left(\boldsymbol{x},\boldsymbol{\delta}_k^i\right)\right),\boldsymbol{y}^{\text{true}}\right)\right)\quad(11)$$

$$\text{s.t.}\,\boldsymbol{\delta}\in\left[m_k,n_k\right]$$

式中，$\boldsymbol{y}^{\text{true}}$代表真实类别，$L$代表分类损失，$F$代表分类输出，$A_k$代表特定的攻击方法，$\alpha$是迭代步长。每步迭代，都要将对样本$\boldsymbol{x}$的扰动$\boldsymbol{\delta}$限定在$\left[m_k,n_k\right]$范围内。随后，用特定的方式对不同语义攻击进行排列组合，实现破坏力更强的组合语义攻击CAA（composite adversarial attack）并最终将其用于对抗训练过程。这种语义攻击与传统的对比度变化等不同，它利用了梯度攻击的思想，并对扰动范围进行了限定，减少扰动值对图像主要信息和人眼感知的干扰，是对抗攻击框架下的语义攻击方法。

注意，这是一种通过在多种扰动的排列组合上对抗训练以获得鲁棒性的方法，看似可见攻击变多，但CAA的攻击目标是为给定的一组攻击找出一个能使其整体破坏力达到最大的排序方式，因此即便攻击种类和数量相同，如果组合攻击顺序不同，最终的组合攻击效果也会不同，在测试时采用这种不同于训练时的组合攻击方式，这种用于测试的攻击对于模型而言仍然是未知的。

12）SSAT。对抗攻击不止在分类任务中存在，车辆轨迹预测任务同样极易受到对抗扰动的影响，但现有工作很少关注防御这种扰动。鲁棒轨迹预测模型的训练面临诸多不同于分类任务的难题：丰富的上下文信息、对抗扰动的随机性、类别标签的缺乏等，其中，所谓的上下文信息是指车辆轨迹可以传达一定的未来车辆行为信息，而这种信息显然是传统对抗防御无法适应的。为解决这些问题，Jiao等人（2023）分析了3种适用于车辆轨迹预测的典型对抗攻击，并由此提出SSAT（semi-supervised semantics-guided adversarial training）方法。在轨迹预测任务中，这3种典型的攻击的示意图如图2所示。其中，实线和虚线分别代表历史轨迹和未来轨迹，点型虚线和线型虚线分别为原始轨迹和对抗轨迹。显然，这种轨迹偏移在现实中很可能是灾难性的。为处理这些不同的偏移，SSAT在对抗训练中按照轨迹偏移的方向对不同特征进行解耦，实现对不同偏移特征的同时处理，从而使模型的防御能够泛化到不可见的轨迹偏移。

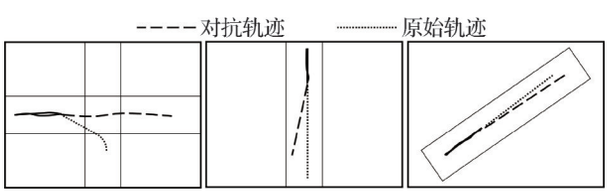

图2　车辆轨迹对抗攻击示意图

Fig. 2　Adversarial attacks of vehicles' trajectories

13）RiFT。Zhu等人（2023）从权重扰动的角度出发对经过对抗训练的模型进行鲁棒性评估，通过评估模型不同模块对鲁棒性的贡献程度，发现在传统对抗训练方法中，仍然存在阻碍模型泛化性的潜在扰动分布。由此，该方法提出了一个鲁棒性的评

价指标 MRC(module robust criticality)，对模型权重进行扰动，观察模型鲁棒性的损失情况，如果损失情况较为微小，则对此权重的调整显然不会严重损害鲁棒性，而对于传统对抗训练方法而言，这部分参数的调整所代表的潜在扰动分布显然没有得到进一步的探索与利用。进一步地，所提出的 RiFT(robustness critical fine-tuning)方法对此进行了针对性的微调，在不损害鲁棒性的前提下，提取出了更多有利于泛化性的鲁棒特征。

相比于上述潜在困难样本的发掘方法，此方法的最大不同之处在于，它没有从样本特征的角度直接探索扰动分布，而是先以模型已经获得的鲁棒性为导向，通过权重扰动的方式定位没有被挖掘的潜在扰动分布，再对这些分布进行学习，避免了面向样本的通用攻击的设计困难。这种权重扰动的方法也为高效的鲁棒微调指明了方向，具体见后文阐述。

14)基于特殊攻击的对抗训练方法总结。在对抗训练中，通过不断地将攻击生成的对抗样本引入训练，模型最终学习到鲁棒的样本特征。因此，对抗训练可看做是对扰动空间的探索、学习过程。进一步地，对抗训练时的攻击对扰动空间的探索能力越强，模型对未知攻击的防御能力自然也越强。这里归纳的特殊攻击方法从不同角度为对抗训练挖掘到更丰富的潜在对抗样本，有效提升了模型对未知攻击的防御能力。为了进一步分析基于特殊攻击的方法的特性，这里对这些方法进行进一步的总结归纳。表 2 对以上方法进行了对比阐述。具体而言，这些方法具有以下几点特征：

1)利用生成器得到更丰富的训练数据。生成器的输入和输出代表了不同的数据分布，在输入层面，可直接添加扰动得到输出的对抗样本，也可将噪声解耦为不同信息，进一步合成得到新的样本；在输出层面，可以直接在生成器的输出上添加扰动。在输入层面添加扰动后，在生成器的影响下，这种扰动在输出层面具有更多的变化，而直接在输出上添加的扰动虽然没有这种变化，但生成器本身的结构特点也会赋予最终结果一定的随机性。由此可知，由生成器产生的训练数据突破了传统扰动的线性干扰方式，赋予了模型以更加发散的方式探索扰动分布的能力。

2)合理拓宽扰动范围。理论上，能够对模型造成严重破坏的对抗攻击方式并不局限于特定范数的传统扰动这一种。除利用生成器增加随机性以外，也可利用模型本身的特征距离放宽扰动值的添加范围，或者利用排列组合方式融合不同扰动。基于此，训练数据覆盖的扰动范围越大，模型对未知攻击的防御能力便越强。

3)直接处理数据分布偏移。数据分布偏移是对抗扰动的常见影响之一。由上述方法可知，对数据分布进行整体扰动虽然较为极端，但是通过对抗训练，模型可以直接学习消除这种整体扰动造成的偏移。此外，领域泛化方法是处理分布偏移的常见工具，可以运用到对抗训练中处理分布偏移。调研发现，利用领域泛化知识实现泛化防御的工作不多，这可能与对抗样本数据分布不稳定等因素有关，但考虑到二者在分布偏移方面的理论联系，这仍然是一个可探索的研究方向。

4)防止过拟合。插值、标签平滑和模型输出的置信度等均与过拟合高度相关。模型通常很难同时探索到所有的扰动区域，因此模型对样本的正确分类能力未必能在这些区域的对抗样本上得以保持。由上述方法可知，相比于促进模型在潜在的对抗样本上学习抵抗正确类别的低置信度，直接接纳这种低置信度，再根据置信度将对抗样本抛弃，可使模型的防御能力高效地延伸到未知区域。此外，传统的插值法在未知攻击的防御中仍然有很大的作用。通过在样本空间、标签空间、损失函数等层面进行插值，以及将插值法与生成器等方法相结合，可以有效防止模型陷入过拟合。这些防止过拟合方法的巧妙运用为针对未知攻击的对抗防御提供了新的思路。

1.1.2 基于正则化约束模型训练

上述方法帮助对抗训练机制更好地探索扰动区域，通过直接修改内部攻击机制来提升防御模型的未知攻击防御能力。相比之下，不在对抗训练流程上做明显改动，而将注意力放在通过正则化提升鲁棒特征质量上的思路吸引了一些研究者的目光。

聚类是提取不同数据共同语义特征的常见方法。Silva 等人(2022)为了减少对抗样本特征的偏移，首先预训练了一个模型，对数据在该模型上的特征进行聚类，得到一个决策模型。随后选择对每一类分别进行聚类，每类数据的聚类中心数目不一定相同。完成聚类学习后，在正式的训练阶段设置一个正则化项，根据聚类决策模型减少对抗样本的特征偏移。这种特征提取方式突破了原始类别的限

表2　基于特殊攻击的对抗训练方法比较

Table 2　Comparison of the adversarial training methods based on well-designed attacks

方法	概述	特征
ADT	最大化样本周围的扰动空间的整体破坏力，通过对抗训练使模型对整体扰动分布均具有鲁棒性	从整体分布的角度探索扰动空间，但是依赖于扰动分布的建模方法
DMAT	对生成器的输入和输出分别进行扰动，鼓励模型通过对抗训练学习基于生成器的样本特征	充分利用了生成器的数据分布建模能力，但是训练数据质量依赖于生成器的质量
CCAT	最大化正确标签之外的所有标签的置信度以生成对抗样本，并对标签进行平滑，防止对抗训练过度专注于正确分类对抗样本	根据扰动程度降低模型对于对抗样本的分类能力，有利于通过置信度识别更广范围的对抗样本，但对模型的原始分类性能造成了损害
PAT	利用网络本身的特征生成对抗样本并加入对抗训练中，力求在符合人眼感知的范围内扩充对抗训练中的扰动空间	弥补了基于范数的扰动范围在人眼感知范围建模上的缺陷，但是扰动距离计算量太大
Lagrangian AT	在攻击中设定扰动惩罚项，放松传统优化攻击的扰动最小化约束，进一步结合动态策略提升可见攻击的扰动丰富性	计算量相对较低，学习到的鲁棒特征不存在偏好问题，但没有脱离梯度、优化等攻击框架
MNG-AC	每轮选定不同的攻击进行对抗训练，通过输出一致化帮助模型学习鲁棒的不变特征	减少多类型扰动同时引入训练的开销，但数据增强在训练过程中的作用有待进一步挖掘
IJSAT	结合插值和生成器提升已知攻击破坏力，缓解对抗训练的自然精度损失和过拟合问题	创新性地通过损失函数插值生成对抗样本，但插值法在生成器方法中的运用有待进一步探索
GAT	同时引入细微噪声和风格扰动，促使模型在对抗训练中学习更高阶的风格特征	充分利用模型结构生成语义信息变化更丰富的可见对抗样本，但无扰动范围限制可能导致训练过程学习到无用特征
NCAT	通过结合扰动值的动态调整策略和不同范数攻击实现模型的跨范数保护	学习过程更加高效，但训练时的扰动变化方式不够多样
REx	通过领域泛化方法提升多类型扰动对抗训练中模型对不同领域数据分布的学习能力	提升了对抗训练对于数据分布偏移的处理能力，但容易受到对抗训练数据分布不稳定的影响
Semantic-GAT	在对抗攻击框架下设计语义扰动，使模型在训练时直接学习这些扰动的高阶排列组合	充分挖掘语义扰动的破坏力，有效提升模型对未知语义攻击的鲁棒性，但未必能保证对非语义攻击具有足够鲁棒性
SSAT	按照车辆轨迹偏移方向对不同偏移特征进行解耦，使模型同时学习不同偏移特征	实现了更高效的未知轨迹攻击防御，但特征解耦的防御思路受到轨迹攻击多样性和模型性能的影响
RiFT	通过权重扰动评估模型不同模块的鲁棒性贡献程度，结合插值方法实现对未知攻击的防御	避免了针对样本的特殊攻击的设计，减小了运算成本，但微调后最终模型参数要基于插值结果仔细权衡

制，对特征空间进行了更加精细的划分，有助于模型学到更加精确的语义特征。但是，这种对每一类分别进行聚类的方法不适用于类别数目较多的大数据集，且由于对抗扰动形式多种多样，这种简单的聚类可能无法完全适应不断变化的扰动。

从统计的角度看，样本在受到攻击时，会以不同的可能性输出不同的错误类别，模型对其输出的次数最多的类别代表了该样本对应的主要攻击方向，而这种方向恰恰很有可能包含重要的对抗样本内在信息。Tack等人(2022)提出了一个正则化项，旨在对不同的数据增强样本进行扰动得到对抗样本，以这些对抗样本的输出一致化为目标进行学习，使模型在学习正确分类的同时把握每个样本的主要攻击方向。直觉上，将这二者结合起来进行优化，使得对抗样本的错误分类概率因正则化而被集中至某个类，又进一步因学习正确分类而被压低，因而学到的特征更加鲁棒。相比于Silva等人(2022)的聚类正则化方法，这种正则化主动接纳对抗样本的错误倾向，从挖掘攻击结果本身的语义信息的角度处理分布偏移，为防御方法提供了新的思路。

1.1.3　基于领域适应提取鲁棒特征

与领域泛化相似，领域适应方法同样可以被引入到未知攻击的泛化性防御中。领域适应不需要过多的领域数目，模型结构相对而言更为直观，因而更易于与对抗防御结合。

直觉上，领域适应与对抗训练的目标是一致的，领域适应旨在使得模型在源域和目标域上的对应输出尽可能一致，而对抗训练旨在使自然样本和对应的对抗样本上的输出尽可能一致。DIAL（domain invariant adversarial learning）（Levi 等，2021）将自然样本作为源域，将对应的对抗样本作为目标域，从而建立起这种理论联系。具体而言，DIAL 的总目标是正确分类自然样本和其对应的对抗样本，同时通过领域标签分类学习它们的不变表示。拉近自然样本和对抗样本的特征以学习不变特征的过程，使得 DIAL 具备泛化到未知攻击的能力，但是这种能力显然依赖于训练阶段所选取的攻击算法。

此外，AFD（adversarial feature desensitization）（Bashivan 等，2021）方法同样注意到了自然样本和对抗样本的分布差异问题，由此引入领域适应来设计对抗训练。AFD 的逻辑与 DIAL 相似。与 DIAL 不同的是，AFD 更偏向于标准的生成对抗思想，即判别器学习区分自然样本和对抗样本，而生成器学习如何迷惑判别器，即最大化对抗样本输入判别器后的输出，学习不变特征。AFD 并没有要求对抗样本的分类预测是否正确，而 DIAL 要求了这一点。除此之外，DIAL 提供了自然样本和对应的对抗样本的分类输出一致性的优化方法，作为正确分类对抗样本的优化目标的替代选项，因此 DIAL 在实现时更加灵活。

Zheng 等人（2024）提出 ADV-4-ADV（adversarial domain adaptation to counter adversarial perturbations），从有利于分类预测的角度对领域适应防御策略进行了分析。该方法的优化过程与前面基本相同，所不同的是，为了保证分类精度，ADV-4-ADV 为每个类别分别分配了一个领域判别器，最大程度地保留每类数据分布的独有特征，实现更加精细的领域特征提取。

1.1.4　基于预训练和微调实现高效鲁棒泛化

自然训练领域的标签收集始终是一个难题，而预训练和微调策略引入了弱监督思想，有效缓解了运算开销。Chen 等人（2020b）尝试将预训练和微调融入对抗训练，提出了若干种可选的训练策略。具体而言，特定预训练任务的优化目标可大体表达为

$$\min_{\theta_p} L_p\left(\theta_p, \theta_{pc}\right) \tag{12}$$

式中，θ_p 和 θ_{pc} 分别是预训练任务的基本参数和额外参数，L_p 是预训练损失。相应地，微调任务的优化目标可大体表达为

$$\min_{\theta} L_f\left(\theta_p, \theta_f\right) \tag{13}$$

式中，θ_f 是分类器的参数，L_f 是分类损失，分类器学习对预训练任务得到的特征进行正确分类。

预训练和微调均可直接融入到标准对抗训练中，具体方法如表 3 和表 4 所示。表中的自然训练策略视为标准对抗训练扰动为 0 的特殊情况。

表 3　预训练任务对抗训练策略

Table 3　Strategies of adversarial training for pre-training tasks

预训练方法	对抗训练损失	对抗训练参数	对抗训练数据集
自然训练	L_p	θ_p, θ_{pc}	预训练数据集
对抗训练	L_p	θ_p, θ_{pc}	预训练数据集

注：两种策略的基础是对抗训练的双层优化过程，这里将自然训练视为双层优化的特殊形式，即内层的扰动值为 0，微调任务同样如此；在预训练阶段，其损失函数形式、数据集种类固定，而预训练任务整体参数均要引入预训练中。

表 4　微调任务对抗训练策略

Table 4　Strategies of adversarial training for fine-tuning tasks

微调方法	对抗训练损失	对抗训练参数	对抗训练数据集
自然训练	L_f	θ_f	微调数据集
对抗训练	L_f	θ_f	微调数据集
自然训练	L_f	θ_p, θ_f	微调数据集
对抗训练	L_f	θ_p, θ_f	微调数据集

注：在微调阶段，其损失函数形式、数据集种类固定，而不同策略的训练参数有所区分：对于自然训练，微调过程须在预训练参数基础上进行，一种是预训练、微调参数均参与训练，另一种是预训练参数固定，仅微调参数参与训练，对于对抗训练同样如此。

上述组合策略减轻了对抗训练的运算开销，充分利用预训练和微调策略在特征学习上的优势，使模型在训练时有机会学到质量更高的隐层特征，从而增强对未知攻击的泛化防御能力，这种结合预训

练和微调的策略为防御研究提供了新的思路。

上述方法为对抗训练在预训练、微调下的策略提供了若干种选择，对这些选择进行了体系化的归纳，展现出预训练、微调策略在鲁棒泛化方面的优越性。然而，上述方法侧重于对抗训练的外部设计。相比之下，Zhu等人(2023)将重点放在了目标模型本身，结合微调策略，提出了RiFT(robustness critical fine-tuning)方法。该方法首先设置了一个模块鲁棒贡献度的评价指标MRC(module robust criticality)，并基于此选择出对于模型鲁棒性而言过剩的模块，在该模块上的微调并不会对模型鲁棒性造成显著影响。由此，RiFT首先在对抗训练模型上通过MRC选出需要微调的模块，接下来在自然样本上完成微调，最后通过微调前后的参数插值选出兼顾鲁棒性和泛化性的参数值。这种方法不仅专注于模型本身的鲁棒性结构问题，而且以一个合理的指标精准指明了微调的方向，进一步发挥出微调策略高效性的优势，可以作为一个加速工具被迁移至多种对抗训练方法中，提升它们对未知攻击的鲁棒性。

1.1.5 基于样本重加权区分样本脆弱性

上述方法均未显式地在训练时区分不同样本。不同的样本在面临不同的攻击时会展现出不同的脆弱性，因此对鲁棒性的贡献也会有所不同。然而，直接根据脆弱程度全局性地为每个样本分配权重并未考虑到样本面对不同攻击情况时的不同脆弱性。因此，若要处理多攻击的情况，就要用更精细的方法分配权重。为此，Gao等人(2021)提出了LRAT(locally reweighted adversarial training)，采用了如下的局部加权方法，具体为

$$\begin{cases}\min\limits_{\theta}\dfrac{1}{N}\sum\limits_{i=1}^{N}\{\alpha_i\ell\left(f_\theta\left(\boldsymbol{x}_i\right),\boldsymbol{y}_i\right)+\\ \quad\sum\limits_{j=1}^{M}\omega\left(\boldsymbol{x}_{ij}^{\mathrm{adv}},\boldsymbol{y}_i\right)\ell\left(f_\theta\left(\boldsymbol{x}_{ij}^{\mathrm{adv}}\right),\boldsymbol{y}_i\right)\}\\ \alpha_i=\max\left(0,\lambda-\sum\limits_{j=1}^{M}\omega\left(\boldsymbol{x}_{ij}^{\mathrm{adv}},\boldsymbol{y}_i\right)\right)\end{cases}\tag{14}$$

式中，N是样本数量，M是每个样本的不同种类对抗样本数，ω是权重，f_θ是目标网络，ℓ是损失函数，$\boldsymbol{y}_i$是自然样本$\boldsymbol{x}_i$正确类别。通过对每个样本对应的对抗样本$\boldsymbol{x}_{ij}^{\mathrm{adv}}$进行加权，实现更精细的权重分配。前一项设置了一个常量λ，避免重加权的对抗训练忽略自然样本。具体而言，每种攻击都有各自的特定加权方式，对于一个自然样本的每一种对抗样本，分别按照其加权规则进行加权，样本被破坏的程度越大，权重越大。如此一来，该方法可在训练中赋予脆弱的样本更高的权重，并用不同的权重分配方式处理不同的攻击，有效避免了脆弱性在不同场景下的相对变化问题，通过更精细的加权方式更加精准地弥补模型的鲁棒性缺陷，从而实现对未知攻击的泛化防御能力。然而，这种能力明显受制于训练时已知攻击的类型和数量。

1.2 以自然训练为中心

1.2.1 基于数据增强提升样本多样性

在自然训练策略下，训练样本的多样性不足是防御模型对未知攻击的鲁棒性较差的重要原因，而数据增强作为一种直观且较为成熟的技术，自然而然地被一些防御工作所采纳，以使防御模型学习到更加丰富的样本特征。

Nandy等人(2020)提出了RBF-CNN(convolutional neural network with the radial basis function)方法，在目标网络前添加一个数据增强模块来建模特定的数据分布，生成更丰富的扰动。具体而言，直接对整个数据分布进行建模过于困难，因此该数据增强模块首先将原始图像分割成若干个块，再根据内容相似度将这些块与一系列滤波核进行匹配。在重建原始图像时，首先对某个块A在所有核上计算匹配分数，形成块A的分数向量，再基于每个核的分布对每个核采样得到一个块，根据之前的分数向量对这些块进行加权求和，形成该块A的重建块。对每个原始块重复上述操作，最终将所有重建块组合形成完整的重建图像。在重建过程中，采样不同的块时可在其中加入一定的噪声，该设计能够进一步控制样本的丰富度，从而影响模型对未知攻击的鲁棒性。

ALAT(attack-less adversarial training)(Ho等，2022)同样追求探索对抗样本的感知边界。现有的防御通常选取若干特定类型的对抗样本进行训练，这些用于训练的对抗样本通常满足人眼难以感知扰动信息的特点。然而，就人眼感知判断的角度而言，难以感知的扰动不会影响人的判断，而即便这些扰动大到可被人眼轻易察觉，只要这种大扰动不过多破坏目标的结构语义信息，人眼依然能够得到正确的判断。这种大扰动的对抗样本是常规对抗样本的极端情况，从感知的角度，它包含几乎所有的常规对抗样本。Ho等人(2022)提出从色彩变化的角度实

现这种大扰动，因为人在识别物体时不会完全关注色彩信息。首先将原图像的像素值范围分割成多个片段，再将每个像素的像素值映射到分割后的某个特定的片段内，在该片段内随机选取一个值作为其最终像素值。这种方法实现了对图像色彩分布的简化，极大丰富了对抗样本的特征。此外，ALAT仅通过一个线性单元实现这种映射，因此即便它被攻击者获知，也可以轻易重新换一个。这为通过数据处理实现鲁棒性的方法提供了一个新思路，即除了花费较高成本保护数据处理模块，也可不进行任何防御，直接简化结构，提升其可替换性，降低部署成本。

1.2.2　基于正则化实现稳定的模型预测

正则化作为一种辅助工具，能够从不同角度提升自然训练策略的性能。在基于自然训练的对抗防御领域，正则化方法的关键在于使模型在自然训练下的原始预测能力能够稳定地延伸到对抗样本上，许多工作以此为基础对防御机制进行了研究。

对抗样本输入模型后，其误差在网络内部传播过程中会被急剧放大，导致模型输出错误的类别。这种现象可以用针对模型损失函数的泰勒展开来解释，具体为

$$L(\boldsymbol{x}+\Delta\boldsymbol{x})=L(\boldsymbol{x})+\frac{\partial L(\boldsymbol{x})}{\partial \boldsymbol{x}}\Delta\boldsymbol{x} \tag{15}$$

式中，L是模型的损失函数，$\Delta\boldsymbol{x}$是样本$\boldsymbol{x}$的扰动。如果损失函数对输入的梯度很大，此时即便$\Delta\boldsymbol{x}$很小，$L(\boldsymbol{x}+\Delta\boldsymbol{x})$也会变大很多，从而导致模型输出错误的结果。这种梯度的存在对于模型而言是灾难性的，若要获得足够强的模型鲁棒性，就必须抑制这种梯度。Yu等人(2018)以此为基础，提出了以下梯度正则化方法，具体为

$$Gradreg(\boldsymbol{x})=-p_t(\boldsymbol{x})+\max_{i\neq t}p_i(\boldsymbol{x}) \tag{16}$$

式中，$p_i(\boldsymbol{x})$是未经softmax处理的针对特定类别的概率输出，t是样本$\boldsymbol{x}$的真实标签。式(16)是对给定样本在错误类别上的最大概率输出与正确类别的输出求差，通过对该差值在对应样本上求梯度以得到最终的梯度正则化项并加入自然训练中。对错误类别的概率输出取最大值可以保证模型在训练时捕捉到最不利于模型正确分类的梯度信息。注意，这种方法在训练阶段未引入任何的对抗样本，额外引入一次反向传播，即为了优化梯度正则化项，在其本身已经完成一阶求导的情况下，在优化过程中对其进行二阶求导，除此之外，该方法无其他任何额外运算开销。此外，基于自然训练进行梯度正则化没有针对模型的其他的任何特殊设计，因此这种方法可以很方便地迁移到其他模型中。

Croce和Hein(2019)发现，同时实现基于无穷范数限制的鲁棒性和基于1范数限制的鲁棒性可以保证实现对其他任何范数的鲁棒性。由此，该方法为自然训练流程添加了一个正则化项，通过用特定方法合并1范数和无穷范数两种范围所覆盖的区域，在训练阶段对样本附近的特定1范数和无穷范数范围进行拓展，以实现对任意的大于1的范数的鲁棒性，这为跨范数的模型保护提供了新的思路。

Jin和Rinard(2020)提出了这样一个假设，即输入数据通常基于一个相对低维的数据空间来生成，不需要过高的维度进行建模。该方法的基本目标是学习一个在给定自然样本周围输出值稳定的网络，显然，这种输出的稳定性依赖于输入数据的分布。结合上述假设，如果能够对数据空间进行更严格的处理，就有可能提升稳定性。该方法以此为出发点，在自然训练流程中加入了对输入数据流形进行处理的正则化项，使模型在学习正确分类的同时提升其输出的稳定性，从而实现更强的未知攻击防御能力。

上述方法没有在鲁棒性的研究中给予可解释性充分的关注。对抗攻击算法对样本的内部特征干扰非常大，而自然样本和其对应的对抗样本由此产生的内部特征差异通常很难被消除。基于这样的想法，Boopathy等人(2020)提出对这种基于可解释性的特征差异进行限制，以实现基于可解释性的鲁棒训练(interpretability-aware robust training)。该方法的正则化表现为自然样本和对抗样本的内层特征差异，可采用类激活映射来作为该特征。该最大差异的产生是伴随着对抗样本生成的，即以该差异的最大化为目标来生成对抗样本。相比于传统的对抗训练，这种处理更高阶的特征差异的方法鼓励模型内部特征保持稳定，进一步提升了模型对未知攻击的泛化防御能力。

1.2.3　基于预训练对抗难以感知的攻击

扰动信息难以被人眼察觉是不同对抗样本的一个共性，在一个合理的、难以被人眼感知的扰动范围内实现鲁棒性成为许多防御机制的目标。但是，在这样的预设范围内，仅对特定类型的扰动实现鲁棒性不符合未知攻击的泛化性需求，有必要对所有在该范围内的扰动都实现鲁棒性。为了满足上述需

求，首先需要设定一个人眼难以感知的扰动边界。Chhabra等人（2021）设计了这样的边界VIB（visual imperceptible bound），具体思想为：给定自然样本，和不同于其类别的样本中距离其最近的样本，以这二者直线距离为度量方式，样本越靠近哪一样本，则其语义信息被视为与其越相似。为使模型对于在预设的VIB范围内的任何扰动都取得较好的鲁棒性，该方法引入随机高斯噪声来干扰自然样本，使模型在训练过程中对该样本的输出与原始自然样本的输出保持一致。在该训练过程中，自然样本的结果由固定的预训练模型得到，对抗样本输入另一个模型，只训练该模型的参数。

VIB的思想与PAT（Laidlaw等，2021）较为相似。与PAT不同的是，VIB直接根据样本在输入空间中与原始样本的距离远近来预设最大感知边界，而PAT使用深度神经网络本身自适应地达到这一点。该方法的设定显然更加简洁直观，但这种度量显然依赖于数据集本身的分布，不一定足够通用。

1.3 以对比学习为中心

1.3.1 基于无监督方法发掘鲁棒特征

对比学习方法是广受欢迎的无监督学习技术，其基本思想是样本特征相似度的学习，而这种特征恰恰是对抗防御所需要的。如果防御模型能够有效提取这种特征，便能通过样本相似度的度量来防御任意攻击，因为通常情况下，对抗攻击不会对样本原有的整体语义信息造成较大破坏。

Kim等人（2020）提出用对比学习思想生成实例级的对抗样本，不使用样本标签。该方法的思想很简单，即直接调用并最大化对比损失生成对抗样本，干扰它的语义信息，随后将这种样本引入最终的优化过程中，最小化对比损失以学习得到最终模型。但是，通过最大化对比损失来制作对抗样本仅仅实现了相似性的消除，但这种干扰没有固定方向，忽略了样本的语义信息。为此，Wahed等人（2022）提出了SwARo（swapping assignments for robust contrastive learning）方法。首先用数据增强创造一系列样本对，再对每一对的第2个样本进行打乱。对于打乱后的样本对，基于以下目标生成每一对中第一个样本的对抗样本：同属一类的样本尽可能相互不相似；不属于同一类的样本尽可能相互相似。如此生成的对抗样本可以达到扰动既具有方向性又能保留语义信息的效果。最后将这种样本引入对抗训练，对抗训练的损失函数采用对比损失，最终学习得到一个鲁棒的特征提取器。后期要完成分类任务，仅需多训练一个分类器即可。样本对打乱示意图如图3所示。

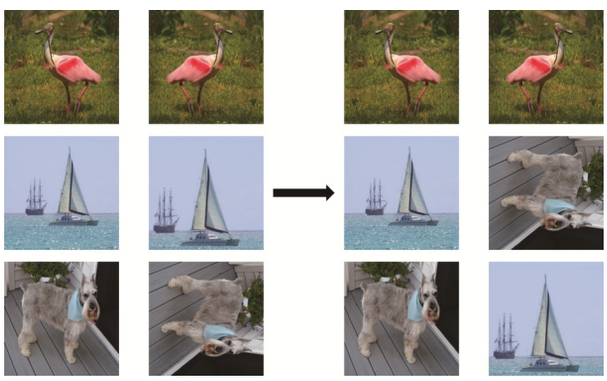

图3 样本对打乱示意图

Fig. 3 Methods of shuffling the example pairs

很多防御模型无法泛化到未知攻击的原因之一是没有学到不同种类对抗样本的鲁棒不变特征。由此，Jiang等人（2020）提出了ACL（adversarial contrastive learning）方法，将对比学习与对抗样本生成结合，同时对齐扰动前后的特征。首先提出了两个对比学习模块：S2S（standard-to-standard），即由不同增强方法得到的自然样本对；A2A（adversarial-to-adversarial），即为增强后的自然样本对添加扰动。最终将S2S与A2A相结合，形成该方法的预训练流程。这4个分支共享所有的卷积层权重，但批归一化层的参数只分别在S2S和A2A内部共享，S2S和A2A之间不共享批归一化层，因为对抗样本和对应的自然样本的统计信息差别很大（Xie等，2020；Xie和Yuille，2020）。在得到预训练的模型后，将其再进行进一步的微调便可得到下游最终模型。

从训练和测试的流程本身来看，很多防御侧重于通过训练来获得鲁棒性，无论其在训练阶段学到的对抗扰动特征泛化性是否足够强，在测试阶段它们都会直接面对攻击算法的破坏。这种在测试时不加任何处理的直接防御仍然暴露出极大的安全漏洞。而生成对抗样本时，图像中的结构信息基本不会受到干扰，因此存在基于结构信息来抵消对抗扰动的可能。Mao等人（2021）提出了一个测试阶段对攻击进行反转的方法。该方法在测试阶段利用对比学习来还原对抗样本，是因为这种方法不关注标签空间，仅关注实例层面的样本相似度，可以基于

相似度自适应地还原不同种类的对抗样本，泛化性强。

1.3.2　基于有监督方法学习鲁棒决策边界

对抗攻击的目标是使模型输出错误的分类结果，即对抗样本的特征偏移至错误的类别。如果将对抗样本的特征从错误类别区域拉回至正确类别区域，则有可能使模型学习到一个更加鲁棒的决策边界。Mao 等人（2019）提出了 TLA（triplet loss adversarial training）方法，设置一个锚点，在将对抗样本拉回正确区域的同时使其远离错误区域。具体而言，TLA 选取对抗样本作为锚点，选取与对抗样本同类别的自然样本作为正例，其他类别的自然样本作为负例。通过拉近锚点和正例，并使锚点和负例互相远离以达到训练目标。

注意，正例和负例均是自然样本，它们在决策空间的位置是稳定的。然而锚点是对抗样本，相比于正例和负例，该对抗样本靠近决策边界的概率更大。以该对抗样本为锚点进行拉近和远离，可使决策边界更容易分类所有样本。如果将上述正例作为锚点，将锚点作为正例，如此进行优化显然没有充分利用决策空间对抗样本的分布特征。

考虑到对比学习方法对样本特征的依赖，计算资源的消耗是以对比学习为中心的防御的重要评价指标。在上述方法中，Kim 等人（2020）的方法和 Wahed 等人（2022）的方法在框架上较为类似，均期望得到一个鲁棒的特征提取器。然而，前者所引入的扰动不具有方向性，且在最后的学习时引入了 3 个对比损失：同一样本的两种增强样本和对抗样本三者之间的相互对比。相比之下，后者的扰动保留了方向性，且最终只引入了自然样本和对抗样本之间的一种对比。因此，后者相对而言比前者消耗的计算资源更少。不同于上述方法未对模型进行分支，ACL 方法（Jiang 等，2020）在实际训练时引入了两个模型分支，因此 ACL 的计算复杂度更易受到模型结构的影响。不同的是，TLA 方法（Mao 等，2019）引入了监督信息，其训练过程在一定程度上避免了对实例信息的依赖。然而，上述这些方法都依赖于通过一定的训练设计防御机制，而 Mao 等人（2021）的测试阶段防御则避免了这个问题，因而其方法的计算资源开销最小。整体而言，考虑到实例层面的丰富信息，如何高效地利用这些信息以减少防御模型的运算开销，是一个值得进一步探索的方向。

2　面向模型架构的未知攻击防御方法

2.1　目标模型结构优化

2.1.1　基于特殊组件弥补模型特征缺陷

深度神经网络在对抗攻击面前的脆弱性与其自身的固有缺陷有必然联系。由此，一些工作专注于寻找模型结构中的脆弱之处，尝试通过修改、替换这些结构等方式来弥补模型的缺陷。

Wen 等人（2020）提出在训练阶段添加有利于模型正确分类的扰动，促使模型在测试阶段利用这种扰动抵消对抗扰动的影响，即在特征层面通过这种扰动将对抗样本拉回正确的决策区域，消除分布偏移。具体而言，将目标网络的全连接层的部分偏置替换为本方法所设计的特殊偏置组件 BPN（beneficial perturbation network），在训练时利用网络损失的梯度直接在特征层生成方向、大小均合理的扰动作为 BPN，消除网络对于对抗样本的错误分类倾向。这种不依赖于输入扰动的训练策略使其可以仅通过在自然样本上训练便可获得一定的鲁棒性，然而这种方法仅在梯度方向上实现了攻击反转，其对非梯度攻击的反转效果仍然值得进一步研究。

Borkar 等人（2020）提出了部分特征重构法（selective feature regeneration）。现有的许多对抗攻击都是依赖于图像和模型信息的攻击算法，如梯度攻击需要利用模型梯度等。一个泛化性足够强的防御应当不仅能防御上面这些攻击，还应当能够防御扰动方式独立于图像的对抗攻击。Borkar 等人（2020）注意到了网络内部特征的脆弱性，并以此为切入点对这种独立于图像的扰动进行防御。具体而言，首先说明了不同卷积核的脆弱性与其权重有关，接下来对一个网络中特定层的不同卷积核按照其脆弱性分成两组，在训练过程中将自然样本和对抗样本混合，以正确分类为目标重新生成选定的比较脆弱的特征，其他特征不参与优化，从而有效弥补了卷积核在提取鲁棒特征的性能上的缺陷。注意，这种对特征的重新生成是基于一个特殊的映射函数组件来完成的，即映射选定特征的输出，满足训练目标。

2.1.2　基于模型分支区分输入分布

模型的结构缺陷有很多种原因，添加特殊组件不是唯一的方法。传统的深度学习模型不加区分地公平处理所有样本，直接得到输出结果。然而在这

种输出模式下，输入数据的不同信息没有被挖掘出来，这为模型在对抗攻击面前的脆弱性埋下了隐患。调研发现，将模型进行分支以显式地利用输入的不同信息同样能够促使模型有效防御未知攻击。

Schott等人(2019)提出的ABS(analysis by synthesis)模型是一个专为MNIST(modified National Institute of Standards and Technology database)数据集(LeCun等，1998)设计的鲁棒模型。即便是MNIST这样简单的数据集，也很少有模型能够在该数据集上实现足够强且足够泛化的鲁棒性。根据特定的物体特征得到相应的类别判断是分类预测的基本因果逻辑，为了使目标模型遵循该逻辑，该方法借助生成器这一强大的数据分布建模工具对每一类的分布分别进行建模，设计了一个贝叶斯分类器。由此，首先引入如下表达式，具体为

$$p(\boldsymbol{y}|\boldsymbol{x})=\frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})} \tag{17}$$

式中，$\boldsymbol{x}$和$\boldsymbol{y}$分别是原始样本和其分类标签。其中，$p(\boldsymbol{y})$可以直接由训练数据估计得到，每个类别的$p(\boldsymbol{x}|\boldsymbol{y})$分别使用一个VAE(variational autoencoder)(Kingma和Welling，2014)估计得到。由于$p(\boldsymbol{x})$固定不变，因此在实际训练中可以忽略此项。ABS没有单独对整个数据分布进行建模，而是对每个类的数据分布分别进行建模，通过贝叶斯思想进行最终分类，它抛弃了传统的深度神经网络分类器，为每一个类别分别引入一个概率输出分支，促使模型通过这种推理式的逻辑进行分类预测，从而避开了对抗攻击的干扰，实现对未知攻击的防御。但ABS是对每个类别分别引入一个VAE，因此它可能无法适应大型数据集。

Xu等人(2021)注意到语义分割任务同样极易受到对抗扰动的干扰，并提出了DDC-AT(dynamic divide-and-conquer adversarial training)。在语义分割任务中，每个像素都有一个分类输出，因此在受到对抗扰动的破坏时，需要对这些像素分开处理。DDC-AT从对抗训练方法出发，在训练目标网络时设置不同的分支来处理脆弱性不同的像素，并动态分配像素所属分支，实现更强的鲁棒性。具体而言，该方法将模型分为3个分支：主分支f_{main}、辅分支f_{aux}和分离分支f_{mask}。首先将给定自然样本中的像素分成A、B两组：A组中的自然样本，其原像素和对抗像素均远离决策边界，二者输出相同；B组中的自然样本，其原像素和对抗像素分属不同类，二者输出不同。将B组输入f_{aux}输出预测值，实现对B组的训练，当B组中的像素变得鲁棒时，便被移到A组，通过f_{main}完成训练。由此，自然样本中的所有像素最终都会被送入f_{main}。对抗样本的像素不用分组，全部输入主分支f_{main}。测试阶段仅使用f_{main}即可，降低运算开销。传统对抗训练对样本所有像素不加区分，统一处理，大大增加了不易被破坏的像素的学习难度，由此降低了自然精度。相比之下，该方法可使模型自适应地处理脆弱性不同的像素，既减少自然精度损失，又提升对未知攻击的鲁棒性。

2.2 输入数据预处理

2.2.1 对抗噪声去除

对抗噪声去除指使用额外的模块去除样本中的噪声，再将去噪后的样本输入至目标网络得到最终结果。去噪方法有着对抗训练无法达到的优势，它无需在训练阶段迭代引入对抗样本，运算开销更小，且对已知噪声的依赖性比对抗训练低，因而更容易泛化到未知噪声。此外，它不需要修改任何模型信息，可迁移性很强。

生成对抗网络(generative adversarial network，GAN)(Goodfellow等，2014)是一种广受欢迎的数据生成方法。Samangouei等人(2018)提出了Defense-GAN方法，即将GAN作为一个预处理模块来对样本进行去噪。首先在原始自然样本上训练得到一个GAN，在测试阶段通过以下方法对样本进行去噪，具体为

$$\min_{z}\left\|G(z)-\boldsymbol{x}\right\|_2^2 \tag{18}$$

式中，G是已经训练好的生成器，$\boldsymbol{x}$是输入样本，z是输入的随机向量。完成上述步骤后，将去噪样本$G(z)$输入目标分类器。由此可见，该方法的基本逻辑是用GAN对原始自然数据分布进行建模，再将对抗样本的分布拉回自然样本的数据分布，以此达到去噪效果。这种方法在训练阶段不依赖任何对抗样本，可以处理任意类型的未知攻击。然而，它完全受限于所使用的生成器的数据分布建模能力，因此仍有许多值得进一步优化的地方。

上述方法没有将重点放在目标模型的特征上。与此不同的是，Xie等人(2019)注意到目标模型深层特征在对抗样本上的脆弱性，即对抗样本在网络的深层特征与其对应的自然样本的特征有非常明显的

差异，由此设计了一个去噪模块，并将其添加进目标网络中，以对抗训练的方法对内含去噪模块的目标网络进行训练，抑制深层特征的噪声，从而提升模型鲁棒性。该去噪模块由特定的滤波核以及残差连接方式（He 等，2016）构成，结构简单且有效。需要注意的是，该方法的去噪模块是嵌入到目标网络中的，它避免了额外去噪模块的设计困难，直接高效地参与到目标网络的对抗训练中。然而，该方法对深层特征的去噪方式只是简单地滤波，而这种原始特征和对抗特征的差异没有被针对性地进行深入挖掘。相比之下，Zhou 等人（2021c）设计了一个去噪模块，以针对性地消除这种差异，具体见下文介绍。

Li 等人（2020）为了提升去噪方法的可迁移性和对大数据集的适应能力，提出了一个间接式去噪的预处理方法 Online Alternate Generator。具体而言，该方法通过合成一个语义信息与输入相似的新图像，以此绕开对抗噪声的干扰。以一个对抗样本为参考样本，引入一个生成器，在训练过程中每优化生成的图像若干次后即优化一次生成器参数，生成器参数优化若干次后停止迭代，输出最终的合成图像，在训练时从数据分布的角度提升合成样本和原始对抗样本的语义相似度。注意，这种方法实用性很强，因为实际部署的目标分类网络的信息一般都经过加密处理，且现实环境中模型将面对更加复杂的场景，对抗训练和其他拓展性较差的方法因而无法被进一步推广。相比之下，该方法不依赖任何攻击算法和目标网络信息，且测试阶段优化参数的策略使得模型参数动态变化，导致攻击者很难通过模型信息制作对抗样本，由此该方法展现出明显的实际应用优势。

对于深度神经网络而言，同一输入中的特征对于模型输出具有不同的重要性。Yang 等人（2021b）提出利用特征解耦的思想设计一个预处理模型 CD-VAE（class-disentangled variational auto- encoder），将输入图像在实例层面的特征根据是否属于分类的关键信息解耦为分类重要信息、分类冗余信息这两种特征。具体而言，CD-VAE 相关分析发现，对抗扰动主要倾向于干扰分类重要特征，而分类冗余特征中仍然对分类有用的信息一般不会被对抗扰动破坏，由此，该方法利用分类重要特征检测对抗样本，用分类冗余特征进行分类预测以提升模型鲁棒性。注意，该方法直接在实例层面进行特征解耦，虽然实例特征相对于内层特征更复杂，但它可以反映出隐层特征无法体现的有关分类的关键特征。

正如 TLA（Mao 等，2019）方法所述，对抗样本通常出现在决策边界处，如果能将其拉回正确的自然数据分布区域，便能够实现鲁棒的对抗防御。Li 等人（2023a）注意到转换函数在对抗防御中的作用，提出了基于转换函数的防御方法 Defense Transformer，即给定一个预训练的分类网络，在训练转换函数过程中，使对抗样本转换后的输出尽可能贴近真实标签。为了提升其对未知攻击的鲁棒性，该方法在训练转换函数时引入多种对抗样本，同时对这些样本进行转换及分类。在转换函数训练过程中，只优化转换函数的参数，分类网络固定不变。

Zhou 等人（2021b）提出了 JATP（joint adversarial training-based pre-processing）方法。现有去噪方法通常只能有效防御由攻击目标网络生成的对抗样本，然而当攻击者可以利用预处理模块的信息攻击预处理模型时，目标模型的鲁棒性显著降低。此外，许多去噪方法只关注对预处理模块去噪能力的提升，没有关注其本身的脆弱性。由此，JATP 专注于对包括预处理模块和目标网络在内的整体模型进行鲁棒防御。具体而言，为防止模型在可见攻击上陷入过拟合，首先通过最大化自然样本和对抗样本的特征距离来对整体模型生成对抗样本，再将其用在预处理模块的训练中，在像素层面进行去噪，在标签和特征层面提升样本分类的正确率。在 JATP 的训练过程中，只有预处理模块参与优化。在标准的去噪框架里，这种整体鲁棒性是去噪方法必须要关注的，它直接关系到模型在实际环境中的性能。

Zhou 等人（2021a）提出了一种基于不变特征提取的去噪方法 ARN（adversarial noise removing network）。首先引入一个编码器提取对抗样本特征，再通过一个判别器判断这些特征的攻击类别，而该编码器学习如何迷惑判别器的判断，即提取攻击不变特征。进一步地，为了从不变特征中还原自然样本，除了利用均方根误差减少像素损失，该方法还引入了另一个图像真假判别器，促使解码器学习如何迷惑该判别器，学习通过不变特征还原得到更加真实的图像。然而，这种方法使其在训练阶段依赖部分类型的对抗样本，而在这些攻击上提取的不变特征未必适应其他差异较大的攻击。为缓解模型的这种特征偏好，该方法引入了一个归一化模块，在学习不变特征的同时，将不同已知攻击的特征分布与高斯

分布对齐，使ARN适应更多的未知攻击。

可解释性鲁棒训练（interpretability-aware robust training）（Boopathy等，2020）表明，深度神经网络的脆弱性与其内部对输入扰动的急剧放大有关。为抑制这种放大效应，Zhou等人（2021c）利用类激活特征设计了一种去噪方法CAFD（class activation feature-based denoiser）。具体而言，类激活特征与输出层直接相连，因而直接影响输出结果，为此该方法首先设计了一种类激活特征对抗攻击CAFA（class activation feature-based attack）：最大化类激活特征距离生成对抗样本，使本攻击对模型输出产生直接而显著的影响。注意，在本攻击中，对于对抗样本的限制只存在于输入像素空间，对类激活特征距离无任何限制，这样可达到样本扰动难以感知但其内层特征被强力破坏的效果。CAFD专注于处理CAFA对抗样本，以最大限度降低样本误差对模型输出的影响。将CAFA对抗样本输入去噪模块，拉近自然样本和去噪样本的类激活特征，再通过传统生成对抗方法提升去噪样本的真实性。相比于先前的可解释性鲁棒训练，该方法不依赖任何样本标签和任何传统攻击方法，且对类激活特征的直接破坏和去噪处理使该方法充分把握住对抗扰动对内层特征破坏的共同特性，提取出有效不变特征，避免了ARN（Zhou等，2021a）中的不变特征偏好问题。

对抗防御发展至今，完美防御所有对抗扰动难度极大。由此，在这种不可避免的复杂威胁下，挖掘对抗扰动的内在信息成为重要的话题。Gong等人（2022）考虑通过对攻击进行反转，即使用RED（reverse engineering of deceptions）思想来达到挖掘扰动内在信息的目的。具体而言，该方法首先考虑了攻击RED的基本形式。具体为

$$\begin{aligned} \boldsymbol{x}_{\text{out}} &= D\left(\boldsymbol{x}^{\text{adv}}\right) \\ \boldsymbol{x}_{\text{out}}^{\text{adv}} &= \boldsymbol{x}^{\text{adv}} - \boldsymbol{x}_{\text{out}} + \boldsymbol{x} \end{aligned} \tag{19}$$

式中，D是反转网络，该方法将其作为去噪器。x是自然样本，$\boldsymbol{x}^{\text{adv}}$是对应的对抗样本，$\boldsymbol{x}_{\text{out}}$是去噪样本，$\boldsymbol{x}_{\text{out}}^{\text{adv}}$是估计得到的对抗样本。然而，该方法通过实验证明，单纯利用重建误差拉近去噪样本与自然样本的距离无法保证模型输出的稳定，由此形成了基于分类判别的去噪方法CDD-RED（class-discriminative denoising-based RED），通过拉近去噪样本和自然样本的像素空间距离，同时使D估计得到的自然样本、对抗样本分别与原始自然样本、原始对抗样本的输出保持一致，最终学习得到一个去噪器D。相比于BPN（Wen等，2020）这种单一扰动的反转方法，CDD-RED通过挖掘并利用内在的扰动信息，进一步改善了网络输出的稳定性。此外，该方法与后续介绍的对抗样本检测有所不同。检测方法倾向于通过样本特征判断一个样本是否为对抗样本，而CDD-RED倾向于对攻击结果背后的机理进行分析诊断，是一种因果推理式的防御，这一点与ABS（Schott等，2019）相似。

Blau等人（2022）引入了近几年非常流行的扩散模型作为去噪模块。首先将样本输入至预处理模块，遵循扩散模型的前向过程加入高斯噪声，再遵循其反向过程对加入一定高斯噪声的样本进行还原，得到预处理后的样本再将其输入分类器。注意，去噪器和目标分类器都在自然样本上训练而成。该方法认为，扩散模型在高斯噪声下可以有效还原图像，而高斯噪声可以有效覆盖对抗噪声，因此扩散模型可以同时去除对抗噪声。然而这种方法依赖于扩散模型前向加噪周期数的选择，即如果选择的周期数过大，则过多的高斯噪声会破坏图像的原有语义信息；如果过小，则达不到理想的去噪效果。因此必须在初始周期和最大周期之间选择一个合理的中间值。Nie等人（2022）提出的DiffPure（diffusion purification）方法也将扩散模型作为预处理模块，去噪方式与前一个方法基本相同，且均存在选择合适的周期数，以在去噪效果和保留结构语义特征之间取得平衡的问题。不同的是，该方法对扩散模型的前向过程在去噪中的作用有这样的理解：前向过程本质是对局部结构的去除过程，因此以对抗样本作为初始输入，在前向加噪过程中，对抗扰动作为一种细微的结构信息会被逐渐平滑掉，从而使对抗样本的分布接近原始样本的分布，与之相关的分析进一步证明了这一点。由此，在DiffPure框架下，其前向和反向过程均可达到去噪效果，而前一个方法倾向于仅将前向过程作为一个数据处理手段得到反向过程的初始化样本，仅将反向过程作为去噪过程。

Xu等人（2022）提出在特征和像素两个空间对自然样本和对抗样本的数据分布进行对齐。具体而言，首先攻击目标网络生成对抗样本，再将对抗样本和自然样本输入生成器得到它们的生成结果。在像素层面，该方法基于像素值差异拉近自然样本和处理后的对抗样本的输入空间分布。在特征层面，对

每一类的自然样本和处理后的对抗样本的分布进行对齐，同时使处理后的对抗样本的每一类的分布更紧密，不同类的分布更分散。此外，需要以目标网络正确分类处理后的样本为目标。最终学习得到的生成器可以对样本进行更加彻底的去噪，提升模型对未知攻击的泛化防御能力。但是，相比于 CAFD（Zhou 等，2021c）等方法，这种去噪方法没有抓住主要的特征信息，方向性不强，优化目标过于复杂。

2.2.2　对抗样本检测

对抗样本检测同样是一种预处理方法。不同于去噪方法的是，它通过利用特定的样本特征区分自然样本和对抗样本，拒绝将对抗样本输入至目标网络，从而提升输出结果的准确性。相比于对抗训练等方法，对抗样本检测技术同样对训练对抗样本的依赖较低，因而具备防御不可见噪声的潜力。

Zheng 和 Hong（2018）注意到，对于一个在自然样本上训练而成的目标网络，部分自然样本被攻击到一个特定的类中，其隐层特征分布与同一类别的自然样本隐层特征分布具有较大差异，由此提出了 I-Defender（intrinsic properties-based defender）方法，利用自然样本在目标网络的内层特征作为检测对抗样本的依据。具体而言，该方法利用混合高斯模型对每一类的分布进行建模，通过设定一个特定的概率阈值来判定输入是否为对抗样本。由此可知，特征分布偏移不仅可以用于鲁棒训练，还可不加处理，直接作为对抗样本检测的依据。

如果样本包含对抗扰动，则其特征变化包含特定的方向信息，如果是自然样本，则其特征没有这种信息。此外，理论上，向对抗样本中加入特定的噪声有可能将其复原为自然样本。Roth 等人（2019）受此启发，提出利用图像中的特定噪声作为探针来找到样本中的某些特定信息。具体而言，向样本加入特定的噪声后，如果其预测结果具有明显的偏向正确类别的倾向，则证明该样本存在能够被这种噪声所抵消的对抗扰动，由此将该样本判定为对抗样本，否则将其判定为自然样本。此方法创新性地利用攻击反转的思想探索扰动内在信息，由此可知样本扰动信息的挖掘是未知攻击防御的重要因素。

Cohen 等人（2020）通过实验发现，对于一个给定的测试自然样本，对其输出影响最大的前 n 个训练样本与训练集中与其特征距离最小的前 n 个样本存在高度的相关性，而测试对抗样本没有这个相关性，由此将这二者结合，提出了 NNIF（adversarial detection using nearest neighbors influence functions）检测方法。对于验证集中的某个特定样本，根据整体训练集对其建立一个 KNN（K-nearest neighbor）模型，再从 KNN 模型中选取对该验证集样本的输出影响最大的前 M 个训练样本的特征向量，由此获取每个验证集样本的这些特征向量。接下来，利用特定的算法生成对应的验证集对抗样本，对此数据集重复上述操作。最终将所有这些特征作为输入，训练一个分类器作为最终的检测器。不同于先前的方法只根据样本本身的信息完成防御，NNIF 创新性地从样本间相互作用的角度挖掘了对抗样本的特征信息，以更通用的视角探索了对抗攻击的作用特性，从而具备更强的未知对抗样本检测能力。

与上述利用样本近邻特征检测对抗样本类似，Abusnaina 等人（2021）所设计的 LNG（latent neighborhood graph）同样采用了这种近邻特征的检测方式，但 LNG 将图神经网络融入到近邻特征的利用中。该方法构建了一个参考样本集合以得到特定样本的 LNG 图，该图由节点信息矩阵和邻接矩阵组成，分别表示图中的节点信息和节点间的连接关系。将这两个矩阵输入至图神经网络中，以此检测该图对应节点是否为对抗样本。该方法中邻接图的信息是可以通过训练来自适应地调整的，节点之间的连接权重与图神经网络参数一同完成训练，这是 NNIF 所不具备的特征提取能力，也凸显了图思想在样本间特征的提取上的优势。

Moayeri 和 Feizi（2021）发现，可以利用自监督思想设计一个简单而高效的分类器 SimCat（SimCLR encoder for catching and categorizing various types of adversarial attacks）检测对抗样本，避免过大的运算开销。该方法通过在一个预训练的自监督编码器的特征上进行线性分类来检测对抗样本。该方法采用预训练的 SimCLR（a simple framework for contrastive learning of visual representations）（Chen 等，2020a）构造检测器，因为实验发现自然样本和对抗样本在该编码器上的隐层特征间的距离与人眼的感知度量水平极为相似。首先利用该编码器获得样本的隐层特征，再将其作为训练样本输入给定的线性分类器，将训练得到的线性分类器作为最终的对抗样本检测器。在其训练过程中，编码器的参数固定不变。注意，SimCat 对未知攻击的泛化防御能力来源于对隐

层特征的选择。这种方法对于扰动距离的度量与LPIPS(Zhang等,2018)较为相似,均以探索符合人眼感知的度量方式为目标。但SimCat有一个显著的优势,即用于度量的特征维度较低,相比之下LPIPS用整个模型的每一层进行度量,运算开销极大。此外,基于SimCat中的距离度量,自然样本和对应的不同种类对抗样本的距离大小相似,这进一步印证了其在感知度量上的优势,并为仅设计一个线性分类器来实现对未知攻击的防御打下基础。

Drenkow等人(2022)从一个极强的泛化视角来审视对抗样本检测方法对已知攻击的依赖问题,认为方法本身应当在尽可能不依赖对抗样本的情况下进行训练,以获得对任意未知对抗样本的检测能力,所提方法RSA(random subspace analysis)以统计特征距离的方式尝试达到这个目标。该方法利用训练样本集在模型特定层的输出得到该层的类别特征,将这些特征与测试样本在该层的特征投影到若干个特征子空间中,通过计算样本与类中心的距离确定样本在每个子空间的标签,最后统计所有子空间的标签一致性,以此检测对抗样本。相比于NNIF和LNG等类似的利用样本间特征的方法,该方法的检测方式更为简单直观,但也明显依赖于网络结构和投影子空间的选取。

Gao等人(2023)引入了互信息方法来设计对抗样本检测器MIAED(mutual information dual autoencoder detector),这是该方法与上述方法的最大不同。MIAED将检测分为两个阶段:首先进行样本特征提取,再进行对抗样本检测。该方法采用了互信息最大化的方式增强特征的可判别性,用分布对齐的方式提升特征的泛化性,并利用一对编码器和解码器提升特征对噪声的鲁棒性,最终提取出高质量的特征。接下来,利用目标网络的输出与该特征的信息进行比较,判断输入是否为对抗样本。由此,第一步得到的特征作为一个参考特征,帮助检测器减少其对目标网络结构和对抗样本的依赖,以实现对未知噪声的防御。然而,与上述基于不同样本相互关系的方法相比,MIAED显然依赖于参考特征的质量,增加了额外的模型设计和部署成本。

上述方法偏向于在训练阶段建立鲁棒的防御机制,这种方式会不可避免地引入一定训练成本,而测试阶段的防御则能有效缓解这一问题。Tsai等人(2023)从测试阶段防御的角度出发,利用掩码自动编码器MAE(masked autoencoder)(He等,2022)建立了一个对抗样本检测器。在该方法中,由于MAE完成了对自然样本分布的精确建模,其在自然样本和对抗样本上的损失函数差异便可作为检测对抗样本的依据。这种不依赖对抗样本,只关注对抗样本与自然样本间的信息差异的方法提升了其对未知攻击的鲁棒性。与SimCat类似的是,这种方法同样只在有效的特征提取方法的基础上进行简单的对比,以此达到检测目标,实现过程简单高效。

3 发展趋势与展望

目前,有关未知对抗攻击的泛化性防御的研究已经取得了一定的进展,呈现为以下几点趋势或问题:

1)测试攻击的不可见性。未知攻击的防御能力直观体现为测试时模型在未知攻击下的输出准确性,以及另外一个非常重要的因素:测试攻击的不可见性。一方面,一些工作仅选取同一范数或同一策略下的攻击作为未知攻击来检验防御能力,这样检验的说服力显然低于跨范数或跨策略防御;另一方面,一些方法在训练阶段不依赖任何攻击,因此任意的对抗攻击对于它们而言都是未知的,这种情况下的未知攻击的不可见性最强,但也最考验防御机制的设计。这种不可见性的考量涉及到模型防御能力、评估标准的选取等方方面面,体现了当下众多防御机制满足复杂场景部署需求的努力,也为将来的研究提供了诸多思考。

2)不变特征的提取。人眼在对图像中的内容做出判断时,习惯于忽略其中的微小扰动,将图像的整体固有特征作为判断的直接依据。由此,模仿这种人类的认知模式以做出判断的模型理论上可以防御任意类型的未知攻击,基于不变特征提取的鲁棒防御也因此广受欢迎。现阶段已有多种提取不变特征的方法相继提出,这些方法涉及到领域适应、领域泛化和对比学习等多个领域。未来,结合多个领域的知识,提取更加通用的不变特征是一个值得进一步研究的方向。

3)对抗扰动的难以感知特性。难以被人眼察觉是对抗扰动的基本特征之一。部分工作尝试在难以感知的扰动范围内对传统攻击进行拓展,力求在该范围内充分暴露深度学习模型的脆弱性,从而使模型学习到的鲁棒特征适用于更多攻击。然而,这种

拓展关联到很多问题。例如，在感知层面，既有工作利用模型本身模拟扰动边界（Laidlaw 等，2021），又有工作采取极端方法攻击图像，使其达到原始语义信息无法被辨认和勉强能够被辨认之间的临界状态，以此作为扰动边界（Ho 等，2022）。但是，上述边界探索方法均未在人眼感知和深度学习模型面对对抗扰动的脆弱性的联系问题上做充分的理论探索，模型的这种脆弱性的根源、人眼在感知层面正确识别对抗样本的生物学原理以及这些因素的相互关联等问题仍然有待研究，因此这些边界的界定方式仍然值得进一步优化。此外，对抗扰动的这些问题与深度学习的可解释性等有必然联系，这同样是未来可行的研究方向。

4）多任务场景下的未知对抗攻击防御。对抗攻击不仅在图像分类任务中存在，也在语义分割、车辆轨迹预测等任务中对模型性能构成了显著威胁（Xu 等，2021；Jiao 等，2023）。在这些新任务中，对抗攻击的机制会有所变化，如车辆轨迹预测中的攻击依赖上下文场景信息，这进一步体现出对抗扰动空间的复杂性。上述方法是针对单一任务设计的防御机制，而现有的一些工作尝试分析不同任务的扰动间的联系，设计一个能够同时处理多个下游任务的泛化性对抗防御机制，减少针对不同任务分别部署防御方法的过大开销，这说明了单一防御方法实现多任务防御在理论上的可行性，也为研究者在设计防御方法时提供了诸多思考（Poursaeed 等，2021）。然而，多任务的扰动必然对模型构成更大、更广泛的威胁，因此面向多任务场景设计泛化性防御机制仍然是当下亟待解决的问题。

5）多模态与大模型技术下的防御。如今，大模型、多模态等人工智能新技术呈现出迅猛发展的态势，展现出强大的应用潜能。然而，这些新兴技术同样面临着恶意攻击的威胁。Gan 等人（2020）抛弃了原始输入空间的攻击方法，转而利用每个模态的深层特征进行对抗训练，将对抗训练方法在多模态学习上进行了扩展。Yang 等人（2021a）发现仅使用单一模态的扰动便能够有效破坏多模态模型的性能，提出对不同模态信息进行一致性检测，并控制不同模态的信息输出，为泛化性更强的多模态鲁棒防御提供了更新的思路。除此以外，大模型技术也面临着恶意攻击的困扰。Xue 等人（2023）设计了一种黑盒攻击框架，通过查询大语言模型的接口生成一种通用的触发器，并设计了一种中毒算法以生成可迁移的中毒提示词，为面向大模型的通用攻击提供了很好的启发。为了防御针对大模型的攻击，Yi 等人（2023）通过基于对抗训练的微调策略帮助模型辨识其外部信息，提升模型对提示词攻击的鲁棒性。此外，Cui 等人（2023）发现大语言模型对于视觉对抗样本的脆弱性，利用特定的文本信息完成大语言模型针对视觉对抗样本的防御，实现了面向跨模态未知攻击的大语言模型鲁棒性。由上述讨论可知，伴随着所要处理的信息范围的扩展，大模型、多模态技术面临着更加多样的恶意攻击的挑战，如跨模态的未知攻击（Cui 等，2023）、迁移性强的恶意提示词（Xue 等，2023）等。此外，相比于传统的深度神经网络，大模型的参数量显然更大，如何使大模型防御方法在保证其鲁棒性的前提下更加高效、通用，是未来值得研究的方向。

除此之外，伴随着应用场景的不断变化，许多与未知对抗攻击的防御相关的其他工作也逐渐发展起来，并取得了不错的研究成果。

1）多类型扰动的鲁棒性。多类型扰动的防御指通过在多种对抗扰动上训练，以同时实现对这些扰动的鲁棒性。这类工作与未知攻击的防御的关键区别在于，前者所防御的扰动均在训练阶段被模型学习过，后者倾向于防御未被引入训练过程的新的攻击。Tramèr 和 Boneh（2019）为对抗训练提出了两种集成式的攻击策略，一种是对于每个样本，在给定的若干种攻击中选择对该样本破坏力最强的来生成对抗样本；另一种是对于每个样本，用给定的所有攻击分别对其进行攻击，再对攻击结果取平均。MSD（multi steepest descent）（Maini 等，2020）对上述集成攻击方法进行了改进，即对给定的若干种攻击，每次迭代从这些攻击中选取破坏力最强的扰动来更新本轮的扰动值，多轮迭代得到最终的对抗样本，这样的集成攻击方法巧妙地以单一攻击的形式集成多种攻击的效果，有效降低了运算开销。Liu 等人（2020）将模型进行分支，每个分支处理一种扰动，提取不同分支的不变特征，从而防御多类型扰动。Wang 等人（2021）从传统对抗训练的双层优化角度出发，通过多领域加权的方式对双层优化方法进行扩展。整体上，这些工作均弥补了传统模型只能实现单一扰动鲁棒性的缺陷，但没有在未知攻击的防御上下功夫。不可否认的是，作为过渡性的工作，多类型扰动的防

御工作在单一扰动防御和未知攻击防御间搭建了一个桥梁，促进了泛化性鲁棒防御研究的进一步发展。

2）未知对抗攻击防御机制的辅助工具。许多辅助性的工作有效促进了未知攻击的防御技术的发展。为了更好地评价未知攻击防御机制的鲁棒性，Kaufmann等人（2019）设计了若干种新的更符合实际场景的攻击，如Fog、Snow等，专门作为测试用的未知攻击，这些攻击已在部分工作中被采用。Sridhar等人（2022）从特征量化、卷积核优化等角度提出了一些鲁棒泛化策略，为未知攻击的防御提供了新的思考。Dolatabadi等人（2022）通过缩减数据集规模来实现高效的鲁棒训练，为未知攻击的防御模型提供了一个加速工具。这些工作进一步满足了未知攻击防御中的高质量评估、模型部署成本缩减等方面的需求。

3）常规未知噪声的泛化性防御。在复杂的现实环境中，对抗扰动并不是深度学习模型的唯一威胁，图像质量很可能经常受到雾气、雨雪、遮挡和光照等多种多样的常规噪声的干扰，它们对模型的破坏不容忽视。调研发现，对这些常规未知噪声的防御同样是当下的研究热点。Hendrycks和Dietterich（2019）创造了专门用于评估常规扰动鲁棒性的数据集CIFAR10-C、CIFAR100-C等，包含了不同类型、不同干扰程度的丰富的常规噪声。Shu等人（2021）在特征层的统计信息上施加扰动以影响特征层对风格信息的编码能力，进一步在这些干扰的特征上训练以获得模型鲁棒性。Song等人（2021）根据攻击轨迹将对抗样本的出现区域进行拓展，提升样本多样性，从而有效防御常规扰动。与此不同的是，Modas等人（2022）设计了一个序列化的数据增强方法来增强样本多样性。Sarkar等人（2022）对样本的不同输出结果进行集合，通过投票机制得到最终的预测结果，以防御常规扰动。综合来看，上述的常规扰动和对抗扰动更加体现了扰动信息的丰富性和现实环境威胁的复杂性。因此，设计一个同时有效防御这两种未知扰动的模型是未来的一个具有重要应用价值的研究方向。

4 结 语

面向未知攻击的对抗防御是当下对抗防御研究的新兴方向。本文系统归纳了未知对抗攻击的防御工作的研究现状，从模型架构和训练策略两个方面对相关工作进行了详细的分析总结。最后对未知攻击的防御工作的相关研究趋势进行了梳理，并对本领域研究的未来发展进行了展望。未知对抗攻击的防御研究对于深度学习可解释性等的研究和模型的更大范围应用等具有重要意义，希望本文可以让更多读者了解本领域的理论与技术，为本领域的进一步发展提供更多有益的思考。

参考文献（References）

Abdel-Hamid O, Mohamed A R, Jiang H, Deng L, Penn G and Yu D. 2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10): 1533-1545 [DOI: 10.1109/TASLP.2014.2339736]

Abusnaina A, Wu Y H, Arora S, Wang Y Z, Wang F, Yang H and Mohaisen D. 2021. Adversarial example detection using latent neighborhood graph//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7687-7696 [DOI: 10.1109/ICCV48922.2021.00759]

Agnihotri S, Jung S and Keuper M. 2023. CosPGD: a unified white-box adversarial attack for pixel-wise prediction tasks [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2302.02213.pdf

Akhtar N and Mian A. 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access, 6: 14410-14430 [DOI: 10.1109/ACCESS.2018.2807385]

Azizmalayeri M and Rohban M H. 2023. Lagrangian objective function leads to improved unforeseen attack generalization. Machine Learning, 112(8): 3003-3031 [DOI: 10.1007/s10994-023-06348-3]

Bashivan P, Bayat R, Ibrahim A, Ahuja K, Faramarzi M, Laleh T, Richards B A and Rish I. 2021. Adversarial feature desensitization//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 10665-10677

Blau T, Ganz R, Kawar B, Bronstein A and Elad M. 2022. Threat model-agnostic adversarial defense using diffusion models [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2207.08089.pdf

Boopathy A, Liu S J, Zhang G Y, Liu C, Chen P Y, Chang S Y and Daniel L. 2020. Proper network interpretability helps adversarial robustness in classification//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: #95

Borkar T, Heide F and Karam L. 2020. Defending against universal attacks through selective feature regeneration//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 709-719 [DOI: 10.1109/CVPR42600.2020.00079]

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//Proceedings of 2017 IEEE Symposium on Security

and Privacy. San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020a. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: #149

Chen T L, Liu S J, Chang S Y, Cheng Y, Amini L and Wang Z Y. 2020b. Adversarial robustness: from self-supervised pre-training to fine-tuning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 699-708 [DOI: 10.1109/CVPR42600.2020.00078]

Chen T L, Zhang Z Y, Liu S J, Chang S Y and Wang Z Y. 2021. Robust overfitting may be mitigated by properly learned smoothening//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Chhabra S, Agarwal A, Singh R and Vatsa M. 2021. Attack agnostic adversarial defense via visual imperceptible bound//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 5302-5309 [DOI: 10.1109/ICPR48806.2021.9412663]

Cohen G, Sapiro G and Giryes R. 2020. Detecting adversarial samples using influence functions and nearest neighbors//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14453-14462 [DOI: 10.1109/CVPR42600.2020.01446]

Croce F and Hein M. 2019. Provable robustness against all adversarial l_p-perturbations for $p \geqslant 1$ [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/1905.11213.pdf

Cui X M, Aparcedo A, Jang Y K and Lim S N. 2023. On the robustness of large multimodal models against image adversarial attacks [EB/OL]. [2024-01-22]. https://arxiv.org/pdf/2312.03777.pdf

Dai L R, Zhang S L and Huang Z Y. 2017. Deep learning for speech recognition: review of state-of-the-arts technologies and prospects. Journal of Data Acquisition and Processing, 32(2): 221-231 (戴礼荣, 张仕良, 黄智颖. 2017. 基于深度学习的语音识别技术现状与展望. 数据采集与处理, 32(2): 221-231) [DOI: 10.16337/j.1004-9037.2017.02.002]

de Jorge Aranda P, Bibi A, Volpi R, Sanyal A, Torr P H S, Rogez G and Dokania P K. 2022. Make some noise: reliable and efficient single-step adversarial training//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: [s.n.]: 12881-12893

Dolatabadi H M, Erfani S and Leckie C. 2022. ℓ_∞-robustness and beyond: unleashing efficient adversarial training//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 467-483 [DOI: 10.1007/978-3-031-20083-0_28]

Dong Y P, Deng Z J, Pang T Y, Zhu J and Su H. 2020. Adversarial distributional training for robust deep learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #693

Drenkow N, Fendley N and Burlina P. 2022. Attack agnostic detection of adversarial examples via random subspace analysis//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 472-482 [DOI: 10.1109/WACV51458.2022.00287]

Gan Z, Chen Y C, Li L J, Zhu C, Cheng Y and Liu J J. 2020. Large-scale adversarial training for vision-and-language representation learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #555

Gao R Z, Liu F, Zhou K W, Niu G, Han B and Cheng J. 2021. Local reweighting for adversarial training [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2106.15776.pdf

Gao S, Wang R X, Wang X X, Yu S, Dong Y Y, Yao S W and Zhou W. 2023. Detecting adversarial examples on deep neural networks with mutual information neural estimation. IEEE Transactions on Dependable and Secure Computing, 20(6): 5168-5181 [DOI: 10.1109/TDSC.2023.3241428]

Gong Y F, Yao Y G, Li Y Z, Zhang Y M, Liu X M, Lin X and Liu S J. 2022. Reverse engineering of imperceptible adversarial image perturbations [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2203.14145.pdf

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial networks [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/1406.2661.pdf

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16000-16009 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He T, Zhang Z, Zhang H, Zhang Z Y, Xie J Y and Li M. 2019. Bag of tricks for image classification with convolutional neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 558-567 [DOI: 10.1109/CVPR.2019.00065]

Hendrycks D and Dietterich T G. 2019. Benchmarking neural network robustness to common corruptions and perturbations//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Ho J, Lee B G and Kang D K. 2022. Attack-less adversarial training for

a robust adversarial defense. Applied Intelligence, 52(4): 4364-4381 [DOI: 10.1007/s10489-021-02523-y]

Hsiung L, Tsai Y Y, Chen P Y and Ho T Y. 2023. Towards compositional adversarial robustness: generalizing adversarial training to composite semantic perturbations//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 24658-24667 [DOI: 10.1109/CVPR52729.2023.02362]

Ibrahim A, Guille-Escuret C, Mitliagkas I, Rish I, Krueger D and Bashivan P. 2022. Towards out-of-distribution adversarial robustness [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2210.03150.pdf

Jiang Z Y, Chen T L, Chen T and Wang Z Y. 2020. Robust pre-training by adversarial contrastive learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1359

Jiao R C, Liu X G, Sato T, Chen Q A and Zhu Q. 2023. Semi-supervised semantics-guided adversarial training for robust trajectory prediction//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: #754 [DOI: 10.1109/ICCV51070.2023.00754]

Jin C and Rinard M. 2020. Manifold regularization for locally stable deep neural networks [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2003.04286.pdf

Jin G Q, Shen S W, Zhang D M, Dai F and Zhang Y D. 2019. APE-GAN: adversarial perturbation elimination with GAN//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 3842-3846 [DOI: 10.1109/ICASSP.2019.8683044]

Kaufmann M, Kang D, Sun Y, Basart S, Yin X W, Mazeika M, Arora A, Dziedzic A, Boenisch F, Brown T, Steinhardt J and Hendrycks D. 2019. Testing robustness against unforeseen adversaries [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/1908.08016.pdf

Kim M, Tack J and Hwang S J. 2020. Adversarial self-supervised contrastive learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #251

Kingma D P and Welling M. 2014. Auto-encoding variational bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: [s.n.]

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Krueger D, Caballero E, Jacobsen J H, Zhang A, Binas J, Zhang D H, Le Priol R and Courville A. 2021. Out-of-distribution generalization via risk extrapolation (REx)//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 5815-5826

Laidlaw C, Singla S and Feizi S. 2021. Perceptual adversarial robustness: defense against unseen threat models//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: OpenReview.net

Lau C P, Liu J, Souri H, Lin W A, Feizi S and Chellappa R. 2023. Interpolated joint space adversarial training for robust and generalizable defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11): 13054-13067 [DOI: 10.1109/TPAMI.2023.3286772]

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

Levi M, Attias I and Kontorovich A. 2021. Domain invariant adversarial learning [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2104.00322.pdf

Li H F, Zeng Y R, Li G B, Lin L and Yu Y Z. 2020. Online alternate generator against adversarial attacks. IEEE Transactions on Image Processing, 29: 9305-9315 [DOI: 10.1109/TIP.2020.3025404]

Li J C, Zhang S H, Cao J Z and Tan M K. 2023a. Learning defense transformations for counterattacking adversarial examples. Neural Networks, 164: 177-185 [DOI: 10.1016/j.neunet.2023.03.008]

Li K C, Wang X Q, Lin H, Li L X, Yang Y Y, Meng C and Gao J. 2022. Survey of one-stage small object detection methods in deep learning. Journal of Frontiers of Computer Science and Technology, 16(1): 41-58 (李科岑, 王晓强, 林浩, 李雷孝, 杨艳艳, 孟闯, 高静. 2022. 深度学习中的单阶段小目标检测方法综述. 计算机科学与探索, 16(1): 41-58) [DOI: 10.3778/j.issn.1673-9418.2110003]

Li Y, Cheng M H, Hsieh C J and Lee T C M. 2022. A review of adversarial attack and defense for classification methods. The American Statistician, 76(4): 329-345 [DOI: 10.1080/00031305.2021.2006781]

Li Z X, Yin B J, Yao T P, Guo J F, Ding S H, Chen S M and Liu C. 2023b. Sibling-attack: rethinking transferable adversarial attacks against face recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 24626-24637 [DOI: 10.1109/CVPR52729.2023.02359]

Liang H S, He E L, Zhao Y Y, Jia Z and Li H. 2022. Adversarial attack and defense: a survey. Electronics, 11(8): #1283 [DOI: 10.3390/electronics11081283]

Liao F Z, Liang M, Dong Y P, Pang T Y, Hu X L and Zhu J. 2018. Defense against adversarial attacks using high-level representation guided denoiser//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1778-1787 [DOI: 10.1109/CVPR.2018.00191]

Lin W A, Lau C P, Levine A, Chellappa R and Feizi S. 2020. Dual manifold adversarial robustness: defense against L_p and non-L_p adversarial attacks//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.:# 294

Liu A S, Tang S Y, Liu X L, Chen X Y, Huang L, Tu Z Z, Song D and

Tao D C. 2020. Towards defending multiple adversarial perturbations via gated batch normalization [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2012.01654v1.pdf

Liu H H, Zuo X Q, Huang H and Wan X. 2022. Saliency map-based local white-box adversarial attack against deep neural networks// Proceedings of the 2nd CAAI International Conference on Artificial Intelligence. Beijing, China: Springer: 3-14 [DOI: 10.1007/978-3-031-20500-2_1]

Lyu H Y, Yu L, Zhou X Y and Deng X. 2021. Review of semi-supervised deep learning image classification methods. Journal of Frontiers of Computer Science and Technology, 15(6): 1038-1048 (吕昊远, 俞璐, 周星宇, 邓祥. 2021. 半监督深度学习图像分类方法研究综述. 计算机科学与探索, 15(6): 1038-1048) [DOI: 10.3778/j.issn.1673-9418.2011020]

Madaan D, Shin J and Hwang S J. 2021. Learning to generate noise for multi-attack robustness//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 7279-7289

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2018. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Maini P, Wong E and Kolter J Z. 2020. Adversarial robustness against the union of multiple perturbation models//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: #616

Mao C Z, Chiquier M, Wang H, Yang J F and Vondrick C. 2021. Adversarial attacks are reversible with natural supervision//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 661-671 [DOI: 10.1109/ICCV48922.2021.00070]

Mao C Z, Zhong Z Y, Yang J F, Vondrick C and Ray B. 2019. Metric learning for adversarial robustness//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #44

Moayeri M and Feizi S. 2021. Sample efficient detection and classification of adversarial attacks via self-supervised embeddings//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7677-7686 [DOI: 10.1109/ICCV48922.2021.00758]

Modas A, Rade R, Ortiz-Jiménez G, Moosavi-Dezfooli S M and Frossard P. 2022. PRIME: a few primitives can boost robustness to common corruptions//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 623-640 [DOI: 10.1007/978-3-031-19806-9_36]

Modas A, Sanchez-Matilla R, Frossard P and Cavallaro A. 2020. Toward robust sensing for autonomous vehicles: an adversarial perspective. IEEE Signal Processing Magazine, 37(4): 14-23 [DOI: 10.1109/MSP.2020.2985363]

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Nandy J, Hsu W and Lee M L. 2020. Approximate manifold defense against multiple adversarial perturbations//Proceedings of 2020 International Joint Conference on Neural Networks. Glasgow, UK: IEEE: 1-8 [DOI: 10.1109/IJCNN48605.2020.9207312]

Nie W L, Guo B, Huang Y J, Xiao C W, Vahdat A and Anandkumar A. 2022. Diffusion models for adversarial purification//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 16805-16827

Poursaeed O, Jiang T X, Yang H, Belongie S and Lim S N. 2021. Robustness and generalization via generative adversarial training// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15711-15720 [DOI: 10.1109/ICCV48922.2021.01542]

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149

Rice L, Wong E and Kolter J Z. 2020. Overfitting in adversarially robust deep learning//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: PMLR: 8093-8104

Roth K, Kilcher Y and Hofmann T. 2019. The odds are odd: a statistical test for detecting adversarial examples//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 5498-5507

Samangouei P, Kabkab M and Chellappa R. 2018. Defense-GAN: protecting classifiers against adversarial attacks using generative models//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Sarkar A, Sarkar A and Balasubramanian V N. 2022. Leveraging test-time consensus prediction for robustness against unseen noise//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1839-1848 [DOI: 10.1109/WACV51458.2022.00362]

Schott L, Rauber J, Bethge M and Brendel W. 2019. Towards the first adversarially robust neural network model on MNIST//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Shu M L, Wu Z X, Goldblum M and Goldstein T. 2021. Encoding robustness to image style via adversarial feature perturbations//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 28042-28053

Silva S H, Das A, Aladdini A and Najafirad P. 2022. Adaptive clustering of robust semantic representations for adversarial image purification on social networks//Proceedings of the 16th International AAAI Conference on Web and Social Media. Atlanta, USA:

AAAI: 968-979 [DOI: 10.1609/icwsm.v16i1.19350]

Song C B, Fan Y B, Yang Y C, Wu B Y, Li Y M, Li Z F and He K. 2021. Regional adversarial training for better robust generalization [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2109.00678v1.pdf

Sridhar K, Dutta S, Kaur R, Weimer J, Sokolsky O and Lee I. 2022. Towards alternative techniques for improving adversarial robustness: analysis of adversarial training at a spectrum of perturbations [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2206.06496.pdf

Sriramanan G, Addepalli S, Baburaj A and Venkatesh Babu R. 2021. Towards efficient and effective adversarial training//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 11821-11833

Sriramanan G, Gor M and Feizi S. 2022. Toward efficient robust training against union of ℓ_p threat models//Proceedings of the 39th International Conference on Machine Learning. New Orleans, USA: PMLR: 25870-25882

Stutz D, Hein M and Schiele B. 2020. Confidence-calibrated adversarial training: generalizing to unseen attacks//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: # 849

Sun C H, Zhang Y G, Wan C Q, Wang Q Z, Li Y, Liu T L, Han B and Tian X M. 2022. Towards lightweight black-box attacks against deep neural networks//Proceedings of the 36th Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 19319-19331

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J and Fergus R. 2014. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: [s.n.]

Tack J, Yu S, Jeong J, Kim M, Hwang S J and Shin J. 2022. Consistency regularization for adversarial robustness//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 8414-8422 [DOI: 10.1609/aaai.v36i8.20817]

Tramèr F and Boneh D. 2019. Adversarial training and robustness for multiple perturbations//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #527

Tsai Y Y, Chao J C, Wen A, Yang Z Y, Mao C Z, Shah T and Yang J F. 2023. Test-time detection and repair of adversarial samples via masked autoencoder [EB/OL]. [2024-01-22]. https://arxiv.org/pdf/2303.12848.pdf

Wahed M, Tabassum A and Lourentzou I. 2022. Adversarial contrastive learning by permuting cluster assignments [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2204.10314.pdf

Wang J K, Zhang T Y, Liu S J, Chen P Y, Xu J C, Fardad M and Li B. 2021. Adversarial attack generation empowered by min-max optimization//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 16020-16033

Wang S and Gong Y X. 2022. Adversarial example detection based on saliency map features. Applied Intelligence, 52(6): 6262-6275 [DOI: 10.1007/s10489-021-02759-8]

Wang X, Li K, Xu M J and Ning C. 2019. Improved remote sensing image classification algorithm based on deep learning. Journal of Computer Applications, 39(2): 382-387 (王鑫, 李可, 徐明君, 宁晨. 2019. 改进的基于深度学习的遥感图像分类算法. 计算机应用, 39(2): 382-387) [DOI: 10.11772/j. issn. 1001-9081. 2018061324]

Wang Z K, Pang T Y, Du C, Lin M, Liu W W and Yan S C. 2023. Better diffusion models further improve adversarial training//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #1507

Wen S X, Rios A and Itti L. 2020. Beneficial perturbations network for defending adversarial examples [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2009.12724.pdf

Weng Z Z, Qin Z J, Tao X M, Pan C K, Liu G Y and Li G Y. 2023. Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Transactions on Wireless Communications, 22(9): 6227-6240 [DOI: 10.1109/TWC.2023.3240969]

Williams P N and Li K. 2023. Black-box sparse adversarial attack via multi-objective optimisation CVPR proceedings//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12291-12301 [DOI: 10.1109/CVPR52729.2023.01183]

Xie C H, Tan M X, Gong B Q, Wang J, Yuille A L and Le Q V. 2020. Adversarial examples improve image recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: #90 [DOI: 10.1109/CVPR42600.2020.00090]

Xie C H, Wu Y X, van der Maaten L, Yuille A L and He K M. 2019. Feature denoising for improving adversarial robustness//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 501-509 [DOI: 10.1109/CVPR.2019.00059]

Xie C H and Yuille A L. 2020. Intriguing properties of adversarial training at scale//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net

Xu X G, Zhao H S and Jia J Y. 2021. Dynamic divide-and-conquer adversarial training for robust semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7486-7495 [DOI: 10.1109/ICCV48922.2021.00739]

Xu X G, Zhao H S, Torr P and Jia J Y. 2022. General adversarial defense against black-box attacks via pixel level and feature level distribution alignments [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2212.05387.pdf

Xue J Q, Zheng M X, Hua T, Shen Y L, Liu Y P, Bölöni L and Lou Q. 2023. TrojLLM: a black-box trojan prompt attack on large language models//Proceedings of the 37th International Conference on Neural

Information Processing Systems. New Orleans, USA: [s.n.]

Yang K, Lin W Y, Barman M, Condessa F and Kolter Z. 2021a. Defending multimodal fusion models against single-source adversaries//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3340-3349 [DOI: 10.1109/CVPR46437.2021.00335]

Yang K W, Zhou T Y, Zhang Y G, Tian X M and Tao D C. 2021b. Class-disentanglement and applications in adversarial detection and defense//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s. l.]: [s. n.]: 16051-16063

Yi J W, Xie Y Q, Zhu B, Kiciman E, Sun G Z, Xie X and Wu F Z. 2023. Benchmarking and defending against indirect prompt injection attacks on large language models [EB/OL]. [2024-01-22]. https://arxiv.org/pdf/2312.14197.pdf

Yin F, Zhang Y, Wu B Y, Feng Y, Zhang J Y, Fan Y B and Yang Y J. 2024. Generalizable black-box adversarial attack with meta learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(3): 1804-1818 [DOI: 10.1109/TPAMI.2022.3194988]

Yoon J, Hwang S J and Lee J. 2021. Adversarial purification with score-based generative models//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 12062-12072

Yu F X, Xu Z R, Wang Y Z, Liu C C and Chen X. 2018. Towards robust training of neural networks by regularizing adversarial gradients [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/1805.09370.pdf

Yuan L, Li X M, Pan Z X, Sun J M and Xiao L. 2022. Review of adversarial examples for object detection. Journal of Image and Graphics, 27(10): 2873-2896 (袁珑, 李秀梅, 潘振雄, 孙军梅, 肖蕾. 2022. 面向目标检测的对抗样本综述. 中国图象图形学报, 27(10): 2873-2896) [DOI: 10.11834/jig.210209]

Zhang B, Zhu J and Su H. 2020. Toward the third generation of artificial intelligence. SCIENTIA SINICA Informationis, 50(9): 1281-1302 (张钹, 朱军, 苏航. 2020. 迈向第三代人工智能. 中国科学: 信息科学, 50(9): 1281-1302 [DOI: 10.1360/SSI-2020-0204]

Zhang H Y, Yu Y D, Jiao J T, Xing E, El Ghaoui L and Jordan M. 2019. Theoretically principled trade-off between robustness and accuracy//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7472-7482

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zheng T Y, Chen Z, Ding S Y, Cai C and Luo J. 2024. Adv-4-Adv: thwarting changing adversarial perturbations via adversar-ial domain adaptation. Neurocomputing, 569: #127114 [DOI: 10.1016/j.neucom.2023.127114]

Zheng X, Fan Y B, Wu B Y, Zhang Y, Wang J and Pan S R. 2023. Robust physical-world attacks on face recognition. Pattern Recognition, 133: #109009 [DOI: 10.1016/j.patcog.2022.109009]

Zheng Z H and Hong P Y. 2018. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 7924-7933

Zhou D W, Liu T L, Han B, Wang N N, Peng C L and Gao X B. 2021a. Towards defending against adversarial examples via attack-invariant features//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 12835-12845

Zhou D W, Wang N N, Gao X B, Han B, Yu J, Wang X Y and Liu T L. 2021b. Improving white-box robustness of pre-processing defenses via joint adversarial training [EB/OL]. [2023-06-07]. https://arxiv.org/pdf/2106.05453.pdf

Zhou D W, Wang N N, Peng C L, Gao X B, Wang X Y, Yu J and Liu T L. 2021c. Removing adversarial noise in class activation feature space//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7878-7887 [DOI: 10.1109/ICCV48922.2021.00778]

Zhu K J, Hu X X, Wang J D, Xie X and Yang G. 2023. Improving generalization of adversarial training via robust critical fine-tuning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4424-4434 [DOI: 10.1109/ICCV51070.2023.00408]

作者简介

周大为，男，博士研究生，主要研究方向为图像处理和对抗攻击与防御。E-mail: dwzhou.xidian@gmail.com

王楠楠，通信作者，男，教授，主要研究方向为计算机视觉和机器学习。E-mail: nnwang@xidian.edu.cn

徐一搏，男，硕士研究生，主要研究方向为计算机视觉、对抗防御。E-mail: ybxu.xidian@gmail.com

刘德成，男，讲师，主要研究方向为计算机视觉、智能安全、图像处理与机器学习。E-mail: dchliu@xidian.edu.cn

彭春蕾，男，副教授，主要研究方向为可视身份可信识别、机器学习、计算机视觉、智能安全和信息内容安全。E-mail: clpeng@xidian.edu.cn

高新波，男，教授，主要研究方向为人工智能、机器学习、图像处理、计算机视觉和模式识别。E-mail: gaoxb@cqupt.edu.cn